Characterization of structure, dynamics and interactions of the lipopolysaccharide-binding protein LBP, the hepatitis B virus X protein (HBx) and the unfolded-state ensemble of ubiquitin

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Leonildo Pedro Cabral Delgado
aus Guinea-Bissau

Basel, 2019

Genehmight von der Philosophisch-Naturwissenscaftlichen Fakultät

auf Antrag von

Prof. Dr. Stephan Grzesiek

Prof. Dr. Timm Maier

Basel, den 20.06.2017

Prof. Dr. Martin Spiess

Dekan

Summary

In this thesis, the structural ensembles and interactions of three proteins in different states have been characterized by high-resolution solution nuclear magnetic resonance (NMR) in combination with several other techniques: (i) the lipopolysaccharide (LPS)-binding protein LBP, which is the first receptor of LPS by the innate immune system, (ii) the hepatitis B virus X protein (HBx), which is involved in hepatitis B virus entry and replication, and (iii) urea-denatured ubiquitin, as a model system for an unfolded state ensemble.

Chapter 1 provides an introduction to the current state of Structural Biology, stating the importance of the three-dimensional structure of biomolecules and of changes in the structure during function. However, it is now recognized that also highly disordered proteins with no fixed three-dimensional structure, such as intrinsically disordered proteins (IDPs), have important functional roles. Related to this phenomenon is the question how a specific amino acid sequence determines the three-dimensional conformation of a protein.

Chapter 2 describes the successful production of the human lipopolysaccharide (LPS)-binding protein (*h*LBP) in insect cells and its biophysical characterization. By circular dichroism (CD) it was shown to contain a mixture of secondary structure elements that is similar to the solved mouse LBP (*m*LBP) structure. Using a growth medium supplemented by isotope-labeled algal extract (AE) and yeast extract (YE), the protein was ^{15}N-labeled and further characterized by NMR. A ^{1}H-^{15}N-TROSY HSQC spectrum allowed to resolve 324 out of 473 expected resonances. In addition, ^{15}N T_1 and T_2 relaxation experiments led to the determination of the rotational correlation time (τ_c) of *h*LBP, verifying its monomeric state in solution. To characterize the interaction of *h*LBP with various glycolipids, two tryptophan residues on the N-terminal tip were used as a probe to assess the binding of glycolipid ligands. Upon binding, the internal quenching of the tryptophan fluorescence was released. This effect was more pronounced for LPS F515 and Lipid IVa than for smaller glycolipids. Additional surface plasmon resonance (SPR) experiments showed that *h*LBP binds to glycolipid aggregates with low micromolar affinity. Further, it was shown by SPR that LBP covers LPS micelles at a 1:3 LBP:LPS stoichiometry. However, only in the presence of the cluster of differentiation 14 (CD14) protein the LPS disaggregates.

Chapter 3 provides a detailed structural and dynamic characterization of the hepatitis B virus X protein (HBx). Using NMR, a nearly complete backbone and side-chain assignment was achieved. The analysis of secondary chemical shifts and ^{15}N relaxation data showed that, despite being highly mobile on the nanosecond time scale, the protein contains four regions with slower backbone dynamics and propensity to form transient secondary structure elements (α-helices and β-strand). Remarkably these regions overlap with known functional motifs of HBx.

Chapter 4 describes a comparison of single-molecule Förster Resonance Energy Transfer (smFRET), NMR and small-angle X-ray scattering (SAXS) data to describe the urea-denatured state of ubiquitin. The analysis showed perfect agreement between the distance-distributions derived from smFRET and NMR/SAXS-restrained ensembles. The NMR measurements supplied important details on local structural propensities and backbone dynamics, and provided even short- and long-range interactions that are only populated to a low percentage. The SAXS data complemented the NMR data with additional constraints of the overall shape of the unfolded-state. The smFRET provided subpopulation-specific distance distributions over a wide range of denaturant concentrations and revealed chain reconfiguration times in the 50-100 ns range. Overall, the combination of the three methods presents the currently most comprehensive description of the structural and dynamic properties of an urea-denatured protein.

Results from this thesis have been or will be published in the following peer-reviewed articles:

Chapter 2: Delgado L, Zähringer U and Grzesiek S (in preparation) Biophysical analysis of the lipopolysaccharide (LPS)-binding protein LBP and its interactions with LPS

Chapter 3: Delgado L, Casagrande F, Bieniossek C and Grzesiek S (in preparation) Structural and dynamic characterization of the intrinsically disordered hepatitis B virus X protein (HBx)

Chapter 4: Aznauryan M, Delgado L, Soranno A, Nettels D, Huang J, Labhardt AM, Grzesiek S and Schuler B Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS *Proc Natl Acad Sci USA* 113(37):E5389–98 (2016)

Acknowledgements

I would like to express my sincere thanks to all the people who directly or indirectly made this thesis possible. First and foremost I would like to express my deep and sincere gratitude to my supervisor Prof. Dr. Stephan Grzesiek, for giving me the opportunity to pursue my thesis in his laboratory and his commitment to educate me on all aspects of science and research from both the theory and practice points of view. I feel highly privileged to have him as my guide.

I could not have wished for a better project than the lipopolysaccharide (LPS)-binding protein LBP, where I had to conduct the research from the beginning to the state described in this thesis. Many people helped me during this process. I would like to extend special thanks to:

Prof. Dr. Ulrich Zähringer, because he was the one who first proposed this project in order to increase the current knowledge on the activation of the innate immune system by LPS. He provided the purest glycolipid compounds for the binding studies described in this thesis. He is also member of my thesis committee, where he always contributed with ideas and suggestions that helped me to drive the project successfully.

Prof. Dr. Timm Maier and his group, especially Dr. Edward Stuttfeld for teaching me so much about X-ray crystallography and for helping me with the screening of crystallization conditions for LBP. In addition, Prof. Dr. Maier is also member of my thesis committee, providing always important inputs on how to overcome the various challenges I faced while conducting my research on LBP.

I would like to thank Dr. Hans-Jürgen Sass for many fruitful discussions about the practical and theoretical aspects of science in general, my thesis projects and life. I also would like to acknowledge him for reading my thesis and providing me his valuable suggestions and corrections.

Marco Rogowski for the preliminary tests and all the technical supported I got when needed.

Dr. Shin Isogai taught me a great deal about research in general, experiment planning and execution and insect cells handling. He was always available for discussions about the project, providing ideas and suggestions.

Dr. Christian Opitz for the exciting contributions on the development of the labeling scheme for proteins produced in insect cells, including LBP. In addition, I could always get input from him on various subjects.

Dr. Larisa Kapinos and Dr. Irena Burmann for their assistance and insights on the SPR studies.

Dr. Timothy Sharpe for his assistance and help using the Biophysics Facility instrumentation to get a better understanding of my biological systems.

I could not go on without thanking all the people involved in the hepatitis B virus X protein (HBx) project:

A special thanks to Hoffmann la Roche team, especially to Dr. Fabio Casagrande and Dr. Cristoph Bieniossenik for proposing the collaboration and for providing the protein for the interesting NMR studies we performed.

Dr. Judith Habazettl for helping me setting up the NMR experiments and with the analysis of the data. Dr. Habazettl was always very supportive and I am indebted for the time she spent helping me.

I would like to express my sincere thanks to all the people involved in the urea-denatured ubiquitin project:

Prof. Dr. Benjamin Schuler and Dr. Mikayel Aznauryan for the collaboration and the beautiful data they produced using single-molecule FRET.

Dr. Alex Labhardt for the ensemble analysis and the time he spent writing the paper. The final article was only possible because of his efforts.

My earnest thanks goes to all colleagues in the group of Prof. Dr. Grzesiek whose intellectual knowledge, both in biology as well spectroscopy, has always boosted me to achieve my goal.

Finally, I would like to thank my friends Hugo Correia and Ricardo Adaixo for the warm welcome even before I moved to Basel and for their continuous support during this period of my life. I would also like to thank my dearest friends Dominik Jedlinkski and Rajesh Sonti for all the adventures we experienced together. It certainly made my PhD days easier.

Leonildo P. C. Delgado

Contents

Summary v

Acknowledgements vii

Contents ix

Abbreviations and symbols xi

1 General introduction 1

Protein structure, flexibility and function 3

Intrinsically disordered proteins (IDPs) 4

Protein folding & Computational approaches 4

Advances in nuclear magnetic resonance spectroscopy 5

Aims of this work 6

References 7

2 Biophysical analysis of the lipopolysaccharide (LPS)-binding protein LBP and its interactions with LPS 11

Abstract 13

Introduction 14

Immunity and pathogen recognition 14

LPS sensing and signaling 15

Role of the LPS-binding protein (LBP) in LPS recognition 17

Results and Discussion 21

Cloning, expression, purification and biophysical characterization of hLBP 21

Large-scale expression and purification 23

Local probing of the LBP-LPS interaction by the intrinsic tryptophan fluorescence 36

LBP-LPS interaction probed by surface plasmon resonance (SPR) 44

Conclusions 55

Materials and Methods 56

Generation of expression construct 56

Preparation of recombinant baculovirus 56

Expression and purification of the insect cell-produced hLBP 57

Gel electrophoresis and Western blotting 57

Size exclusion chromatography and static light scattering 58

CD spectroscopy 58

NMR spectroscopy 58

Preparation of glycolipid samples 58

Surface plasmon resonance 59

Tryptophan fluorescence measurements 60
Mass spectrometry 60
References **61**

3 Structural and dynamic characterization of the intrinsically disordered hepatitis B virus X protein (HBx) **67**
Abstract **69**
Introduction **69**
Results and Discussion **72**
Backbone and side chain resonance assignments 72
Secondary structure propensities 75
Backbone dynamics 77
Disorder prediction from primary sequence in comparison to experimental data 79
Conclusions **82**
Materials and methods **83**
HBx(9CS) construct 83
Expression and purification 83
NMR experiments 83
NMR Resonance Assignments 84
NMR Relaxation Experiments and Analysis 84
References **85**
Supplementary Information **89**

4 Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS **101**
Introduction **103**
References **105**

Abbreviations and symbols

Å	0.1 nm
AE	algal extract
BEVS	Baculovirus Expression Vector System
BMRB	Biological Magnetic Resonance Data Bank
BPI	bactericidal/permeability increasing protein
CAPRI	Critical Assessment of Predicted Interactions
CASP	Computational Assessment of Structure Prediction
CBB	coomassie brilliant blue
CD	circular dichroism
CD14	cluster of differentiation 14 protein
CETP	cholesterol ester transfer protein
CMC	critical micelle concentration
CV	column volume
DDB1	damage-specific DNA binding protein 1
DHPC	1,2-diheptanoyl-sn-glycerol-3-phosphocholine
E. coli	Escherichia coli
FOS-12	foscholine 12
FRET	Förster resonance energy transfer
HBM	honeybee melittin
HBV	hepatitis B virus
HBx	hepatitis B virus X protein
HCC	hepatocellular carcinoma
HSQC	heteronuclear single quantum coherence
Hz	s^{-1}
IDP	intrinsically disordered protein
IEX	ion exchange

IMAC	immobilized metal affinity chromatography
INEPT	insensitive nuclei enhanced by polarization transfer
K_d	dissociation constant
k_{off}	dissociation rate
k_{on}	association rate
LBP	lipopolysaccharide-binding protein
LPS	lipopolysaccharide
MALS	multi-angle light scattering
MD	molecular dynamics
MD2	myeloid differentiation factor 2 protein
MS	mass spectrometry
MW	molecular weight
MWCO	molecular weight cut-off
NMR	nuclear magnetic resonance
NOE	nuclear overhauser effect
NOESY	nuclear overhauser spectroscopy
ORF	open reading frame
PAGE	polyacrylamide gel electrophoresis
PAMP	pathogen associated molecular pattern
PDB	RCSB Protein Data Bank
PLTP	phospholipid ester transfer protein
PTM	post-translational modification
R_{max}	equilibrium surface plasmon resonance response
SAXS	small-angle X-ray scattering
SEC	size-exclusion chromatography
SDS	sodium dodecyl sulphate
smFRET	single-molecule Förster resonance energy transfer

SPR	surface plasmon resonance
T_1	longitudinal relaxation time
T_2	transversal relaxation time
τ_c	rotational correlation time
TEN	triethylamine
TLR4	toll-like receptor 4
TROSY	transverse-relaxation optimized spectroscopy
Trp	tryptophan
UPS	ubiquitin proteasome system
ν_N	larmor frequency of the ^{15}N nuclei
YE	yeast extract

1 General introduction

Protein structure, flexibility and function

Biological processes depend largely on proteins, nucleic acids and the interplay between these macromolecules. A hundred years ago, we knew very little about the structure of these biological macromolecules. In the 1950s and early 1960s, crystal structures provided for the first time atomic details of DNA (1) and the proteins myoglobin (2) and hemoglobin (3).

Since that time, 130102 entries have been deposited in the Protein Data Bank (PDB), of which 89.5 % were determined by X-ray crystallography, 9.1 % by nuclear magnetic resonance (NMR), and 1.2 % by electron microscopy (EM). New structures are solved at an ever increasing rate and higher-resolution, which provide now an almost complete overview of the space of possible protein folds.

However, the structure of a biological macromolecule is by no means static – indeed, it almost always must undergo conformational changes while performing its function. Thus, defining the molecular structure and its dynamic behavior is essential for a complete understanding of the biological function. Recent years have witnessed significant progress in this area. Important fields are: protein folding (4), signal transduction (1, 5), catalytic cycles of enzymes (2, 6), and even the operation of molecular machines and motors (3, 7).

A specific example of a relevant biological system where the function has been understood from the combination of structural, dynamical and biochemical data is the following:

Protein kinases play an essential role in virtually all cellular processes. The very first atomic view on any protein kinase was provided in 1991 by the crystal structure of the catalytic subunit of the protein kinase A (PKA) (8). It revealed structural features that are conserved for all protein kinases. Furthermore, it allowed to explain substrate specificity and highlighted the importance of conformational flexibility. Subsequent structural analysis of PKA with both catalytic and regulatory subunits provided a molecular explanation for the inhibition of PKA and allowed to build a model for cAMP-mediated activation through binding to the regulatory subunit (9, 10). Recently, a 2.3 Å X-ray structure of the intact, tetrameric PKA holoenzyme yielded insights into allosteric regulation, which has ramifications for understanding the regulation of other conserved kinases (11). In the field of allostery, important insights were also obtained by NMR, where Masterson et al. mapped the allosteric network in the catalytic subunit and showed that positive allosteric cooperativity is generated by nucleotide and substrate binding and lead to transitions between the major conformational states (apo, intermediate and closed) (12).

Intrinsically disordered proteins (IDPs)

For a long time the textbook knowledge was that a unique amino acid sequence encodes a unique energetically stable three-dimensional fold associated with some conformational fluctuations that allow for the biological function. In the last two decades, however, several studies have shown that many functional proteins do not always have a unique 3D structure under functional conditions (13).

Such intrinsically disordered proteins (IDPs) have low sequence complexity with a low proportion of bulky hydrophobic amino acids and a high proportion of charged amino acids (14). Despite being functional, they are unable to fold spontaneously into stable, well-defined, globular three-dimensional structures. Instead, they are dynamically disordered and fluctuate rapidly through a range of conformations, spanning from extended statistical coils to collapsed globules (15).

Bioinformatics studies of the complete genome revealed that regions of disorder are very common in eukaryotic proteins (14, 16). It is now known that IDPs have a central role in the regulation of signaling pathways and crucial cellular processes, including the regulation of transcription, translation and the cell cycle (14, 17-19). The level of IDPs in the cell is tightly controlled to ensure precise signaling and deviations from the norm are associated with several diseases (20-22).

Protein folding & Computational approaches

In contrast to IDPs, most other proteins must adopt a specific 3D structure to be functional. As shown by Anfinsen in his seminal experiment (23), all the information necessary to determine the three-dimensional fold of a protein is incorporated in its amino acid sequence. However, the sequence of events to reach a stable conformation from the primary sequence - "the protein folding problem", is still unsolved (24).

As conclusion of the "Levinthal's paradox" (25), folding must take place through predetermined pathways (26). This statement was based on the fact that there are too many possible conformations for a protein to find the "needle" (the native structure) in the "haystack" (conformational space), by random search in the experimentally observed subsecond time scale (27).

To get insights into the protein folding problem, computational approaches, able to generate high-resolution structural models from primary sequences, are becoming more and more important. Besides homology modeling, which can achieve accuracies of 1-2 Å in favorable cases of homologies higher than 30 %, the most successful *de novo* folding method is based on the Rosetta approach, which assembles protein structures from small fragments of local

structural homologues (28-30). For small proteins (< ~100 residues), these methods can yield accuracies of <1.5 Å.

An important breakthrough for the understanding of folding has been achieved by using molecular dynamics (MD) simulations of millisecond trajectories, though which the folding of 12 small proteins with accuracies of about 2 Å could be reproduced (31). This development is remarkable, since it allows very stringent tests and improvements of MD force fields by comparison to the observed micro- to millisecond range dynamics by NMR (32).

Characterizing protein interactions and higher order protein assemblies is another crucial part for the complete understanding of the protein function. Prediction of molecular complexes is, however, difficult even when the structures of the interacting partners are known. This is evident from the recent first joint Computational Assessment of Structure Prediction (CASP)-Critical assessment of Predicted Interactions (CAPRI) competition (33). Results show that the prediction of homodimer assemblies by homology modeling techniques and docking calculations can be successful for targets featuring large subunit interfaces. The prediction is, however, very poor for targets where a smaller pair-wise interface severely limits the ability to derive the correct structure (33).

Advances in nuclear magnetic resonance spectroscopy

Despite all the advancements in predicting the protein structure and interactions, experimental information is still of utmost importance to guide the search for a rational connection between sequence, structure, dynamics and function. In this context, NMR studies play an important role since the method allows to investigate both structural and dynamical information at atomic resolution with minimal invasiveness under nearly physiological conditions. Several progresses in solution NMR methods have extended the range of applications in recent years: the use of methyl-TROSY along with optimized labeling methods to characterize large proteins up to 2.5 MDa (34); the detection of conformational exchange dynamics between protein substates using relaxation dispersion techniques (35, 36); the improved description of the structural distributions of unfolded states (37); and new strategies for detecting protein NMR signals in cells (38). Also solid-state NMR becomes more and more important since it can provide structures of non-soluble, non-crystalline proteins such as amyloid fibrils (39-42) and membrane proteins in lipid bilayers (43, 44).

Aims of this work

Using the described approaches, in particular in the field of NMR, we aimed to increase the knowledge on the structure and function relationships of three proteins:

1 The lipopolysaccharide (LPS)-binding protein LBP, which is the first LPS receptor of the innate immune system, in order to gain insight in the interaction with the bacterial LPS.

2 The hepatitis B virus X protein (HBx), an intrinsically disordered protein involved in hepatitis B virus entry and replication, in order to improve the understanding of its interactions with cellular partners.

3 The urea-denatured ubiquitin, an otherwise well understood, stable folded protein, by the combined use of NMR, single-molecule FRET and computational techniques to obtain a complete quantitative description of its unfolded state.

References

1. Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171(4356):737–738.

2. Kendrew JC, et al. (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* 181(4610):662–666.

3. Perutz MF, et al. (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. *Nature* 185(4711):416–422.

4. Korzhnev DM, et al. (2004) Low-populated folding intermediates of Fyn SH3 characterized by relaxation dispersion NMR. *Nature* 430(6999):586–590.

5. Gardino AK, et al. (2009) Transient non-native hydrogen bonds promote activation of a signaling protein. *Cell* 139(6):1109–1118.

6. Abrahams JP, Leslie AG, Lutter R, Walker JE (1994) Structure at 2.8 A resolution of F1-ATPase from bovine heart mitochondria. *Nature* 370(6491):621–628.

7. Noji H, Yasuda R, Yoshida M, Kinosita K (1997) Direct observation of the rotation of F1-ATPase. *Nature* 386(6622):299–302.

8. Knighton DR, et al. (1991) Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* 253(5018):407–414.

9. Kim C, Xuong N-H, Taylor SS (2005) Crystal structure of a complex between the catalytic and regulatory (RIalpha) subunits of PKA. *Science* 307(5710):690–696.

10. Wu J, Brown SHJ, Daake von S, Taylor SS (2007) PKA type IIalpha holoenzyme reveals a combinatorial strategy for isoform diversity. *Science* 318(5848):274–279.

11. Zhang P, et al. (2012) Structure and allostery of the PKA RIIβ tetrameric holoenzyme. *Science* 335(6069):712–716.

12. Masterson LR, Mascioni A, Traaseth NJ, Taylor SS, Veglia G (2008) Allosteric cooperativity in protein kinase A. *Proc Natl Acad Sci USA* 105(2):506–511.

13. Uversky VN (2014) Introduction to intrinsically disordered proteins (IDPs). *Chem Rev* 114(13):6557–6560.

14. Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* 293(2):321–331.

15. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6(3):197–208.

16. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z (2002) Intrinsic disorder and protein function. *Biochemistry* 41(21):6573–6582.

17. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* 323(3):573–584.

18. Liu J, et al. (2006) Intrinsic disorder in transcription factors. *Biochemistry* 45(22):6873–6888.

19. Galea CA, Wang Y, Sivakolundu SG, Kriwacki RW (2008) Regulation of cell division

by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. *Biochemistry* 47(29):7598–7609.

20. Gsponer J, Futschik ME, Teichmann SA, Babu MM (2008) Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. *Science* 322(5906):1365–1368.

21. Vavouri T, Semple JI, Garcia-Verdugo R, Lehner B (2009) Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell* 138(1):198–208.

22. Babu MM, van der Lee R, de Groot NS, Gsponer J (2011) Intrinsically disordered proteins: regulation and disease. *Curr Opin Struct Biol* 21(3):432–440.

23. Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* 181(4096):223–230.

24. Creighton TE (1990) Protein folding. *Biochem J* 270(1):1–16.

25. Levinthal C (1968) Are there pathways for protein folding. *J Chim Phys* (65):44-45.

26. Englander SW, Mayne L (2014) The nature of protein folding pathways. *Proc Natl Acad Sci USA* 111(45):15873–15880.

27. Dill KA, Chan HS (1997) From Levinthal to pathways to funnels. *Nat Struct Biol* 4(1):10–19.

28. Xu D, Zhang Y (2013) Ab Initio structure prediction for Escherichia coli: towards genome-wide protein structure modeling and fold assignment. *Sci Rep* 3:1895.

29. Zhang W, et al. (2016) Integration of QUARK and I-TASSER for Ab Initio Protein Structure Prediction in CASP11. *Proteins* 84 (Suppl 1):76–86.

30. Ovchinnikov S, Kim DE, Wang R (2015) Improved de novo structure prediction in CASP11 by incorporating Co evolution information into rosetta. *Proteins* 84 (Suppl 1):67-75.

31. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE (2011) How Fast-Folding Proteins Fold. *Science* 334(6055):517–520.

32. Lange OF, van der Spoel D, de Groot BL (2010) Scrutinizing molecular mechanics force fields on the submicrosecond timescale with NMR data. *Biophys J* 99(2):647-655.

33. Lensink MF, Velankar S (2016) Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins* 84 (Suppl 1):323-48.

34. Velyvis A, Yang YR (2007) A solution NMR study showing that active site ligands and nucleotides directly perturb the allosteric equilibrium in aspartate transcarbamoylase. *Proc Natl Acad Sci USA* 104(21):8815-20.

35. Franco R, Gil-Caballero S, Ayala I, Favier A, Brutscher B (2017) Probing Conformational Exchange Dynamics in a Short-Lived Protein Folding Intermediate by Real-Time Relaxation-Dispersion NMR. *J Am Chem Soc* 139(3):1065–1068.

36. Neudecker P, Lundström P, Kay LE (2009) Relaxation dispersion NMR spectroscopy as a tool for detailed studies of protein folding. *Biophys J* 96(6):2045–2054.

37. Aznauryan M, et al. (2016) Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS. *Proc Natl Acad Sci USA* 113(37):E5389–98.

38. Luchinat E, Banci L (2017) In-cell NMR: a topical review. *IUCrJ* 4(Pt 2):108–118.

39. Iwata K, et al. (2006) 3D structure of amyloid protofilaments of beta2-microglobulin fragment probed by solid-state NMR. *Proc Natl Acad Sci USA* 103(48):18119–18124.

40. Jaroniec CP, et al. (2004) High-resolution molecular structure of a peptide in an amyloid fibril determined by magic angle spinning NMR spectroscopy. *Proc Natl Acad Sci USA* 101(3):711–716.

41. Nielsen JT, et al. (2009) Unique identification of supramolecular structures in amyloid fibrils by solid-state NMR spectroscopy. *Angew Chem Int Ed Engl* 48(12):2118–2121.

42. Wasmer C, et al. (2008) Amyloid fibrils of the HET-s(218-289) prion form a beta solenoid with a triangular hydrophobic core. *Science* 319(5869):1523–1526.

43. Park SH, et al. (2012) Structure of the chemokine receptor CXCR1 in phospholipid bilayers. *Nature* 253:1278.

44. Das BB, et al. (2012) Structure determination of a membrane protein in proteoliposomes. *J Am Chem Soc* 134(4):2047–2056.

2 Biophysical analysis of the lipopolysaccharide (LPS)-binding protein LBP and its interactions with LPS

Abstract

Lipopolysaccharide (LPS), the major component of the outer membrane of Gram-negative bacteria, is one of the most potent stimuli of the human immune system. Transfer of LPS to the toll-like receptor 4 (TLR4)-myeloid differentiation factor 2 (MD2) complex is catalyzed by the LPS-binding protein (LBP) in combination with the cluster of differentiation 14 (CD14) protein. While a crystal structure of LPS-TLR4-MD2 exists, the binding of LPS to CD14 and LBP is less understood.

In this study, the expression of the human LBP (*h*LBP) using the baculovirus/insect cell system was successfully established and yielded, routinely, 2-2.5 mg of pure protein per liter of culture. The recombinant protein was extensively characterized by various biophysical methods and shown to possess seven glycosylation sites, with three of them overlapping with the predicted positions found in UniprotKB. Circular dichroism (CD) measurements showed that the *h*LBP is folded and contains a mixture of helices and strands in similar amount as described for the mouse LBP (*m*LBP).

For heteronuclear nuclear magnetic resonance (NMR), a labeling scheme based on a previously established strategy using isotope-labeled yeast extract (YE) was developed, which allowed the incorporation of about 80 % of ^{15}N nuclei in the LBP. Thus, it was possible to record a ^{1}H-^{15}N-TROSY HSQC, where a total of 324 out of 473 expected backbone resonances were observable. Furthermore, the measurement of ^{15}N T_1 and T_2 relaxation experiments allowed the determination of a rotational correlation time of 22.8 ns for LBP, an expected value for a monomeric protein of approximately 50 kDa.

The intrinsic fluorescence emission of LBP increases upon binding to different glycolipids, which indicates conformational changes in the vicinity of the LBP's tryptophans. Comparison of the effects upon binding of different compounds indicated that heparin disaccharide does not bind, likely due to the lack of acyl chains and that Lipid X is the weakest binder due to its reduced number of acyl chains. The importance of the electrostatic interaction was shown, as binding is significantly reduced when the concentration of NaCl in the binding buffer changes from 0.2-2 M.

By SPR, the binding affinity of LBP to the different glycolipids was determined to be in the low micromolar range, with GlcN(3-OH-14:0)-α-P and Lipid X showing the lowest affinity. Analysis of the binding stoichiometry revealed that LBP can bind LPS in a ratio of approximately 1:3, thus suggesting that the protein is capable of coating glycolipid surfaces. The disaggregation of LPS micelles, however, is a complex process that requires the presence of both LBP and CD14.

Introduction

Immunity and pathogen recognition

Humans and other multicellular organisms are continuously challenged by the threat of invading microbial pathogens. The ability to overcome infection depends in part on the adaptative immune system, which generates immunological memory of previous infections in order to mount a specific response. Adaptative immune responses, however, are slow to develop on the first encounter with a new pathogen, as specific B and T cells have to be activated, which can take 4-7 days before the responses are effective (1). This late reaction can be disastrous considering that typical doubling times for bacteria of few hours (e.g. *Salmonella in vivo*) leads to a progeny on the order of 10^8, corresponding to a fully developed infection, in a single day. Therefore, during the first critical hours and days of exposure to a new pathogen, the innate immune system is active, providing the first line of defense.

The innate immune system has evolved proper mechanisms to sense specific arrangements of key molecules known as pathogen associated molecular patterns (PAMPs) (2) of invading pathogens, including fungi, viruses and bacteria. Well-known PAMPs are the bacterial lipopolysaccharide (LPS), peptidoglycans, flagellin, fungal polysaccharides and viral nucleic acids (3). The swift response carried out to eliminate the infection relies on the recognition of both microbes and their products by the pattern recognition receptors (PRRs) (4). Currently, in humans, ten toll-like receptors (TLRs) with different ligand-binding specificities have been identified. These are type I transmembrane proteins present on the plasma membrane (TLR1, TLR2, TLR4, TLR5, TLR6 and TLR11) or within the endosomal and lysosomal compartment (TLR3, TLR7, TLR8, TLR9 and TLR10) (5). While the extracellular domain, containing varying leucine-rich-repeat (LRR) motifs, is involved in the recognition of PAMPs, the intracellular TIR (Toll/IL-1R) domain activates signaling pathways; this leads to the induction of inflammatory mediators such as cytokines and chemokines essential in host defense (6). While the inflammation originating from innate and adaptative immunity is a normal response to infection, exaggerated inflammatory response is harmful to the host (7). Thus, a proper response level is necessary to preserve the host-pathogen homeostasis in order to avoid a prolonged or exaggerated activation of these responses.

LPS sensing and signaling

LPS, or endotoxin, is a ubiquitous component of the outer-membrane of gram-negative bacteria, being indispensable for growth and survival as it confers resistance to hydrophobic and hydrophilic antimicrobial compounds (8). Due to its external location, LPS participates in important interactions with other biological systems, especially in host-bacteria interactions. Once LPS is released from the infecting organism by bacterial lysis or other mechanisms, it plays an important role in the activation of both the innate and adaptative immune systems (9-11). The presentation of LPS associated with the bacterial surface is also capable of eliciting such responses, albeit with significantly less potency than that observed for soluble LPS (12).

LPS is composed of three distinct regions covalently linked to each other (Figure 1): a glycolipid region named lipid A, a core oligosaccharide and the O-specific polysaccharide chain. The core region contains at least one residue of 3-deoxy-D-*manno*-oct-2-ulosonic acid (Kdo), which is a marker of LPS. Lipid A is inserted into the outer leaflet and anchors the LPS in the outer membrane (OM) with the carbohydrate chain oriented outwards. The complete LPS comprising all three regions is known as S-form (smooth form) whereas that in mutants lacking the glycan is called R-form (rough form) LPS (13).

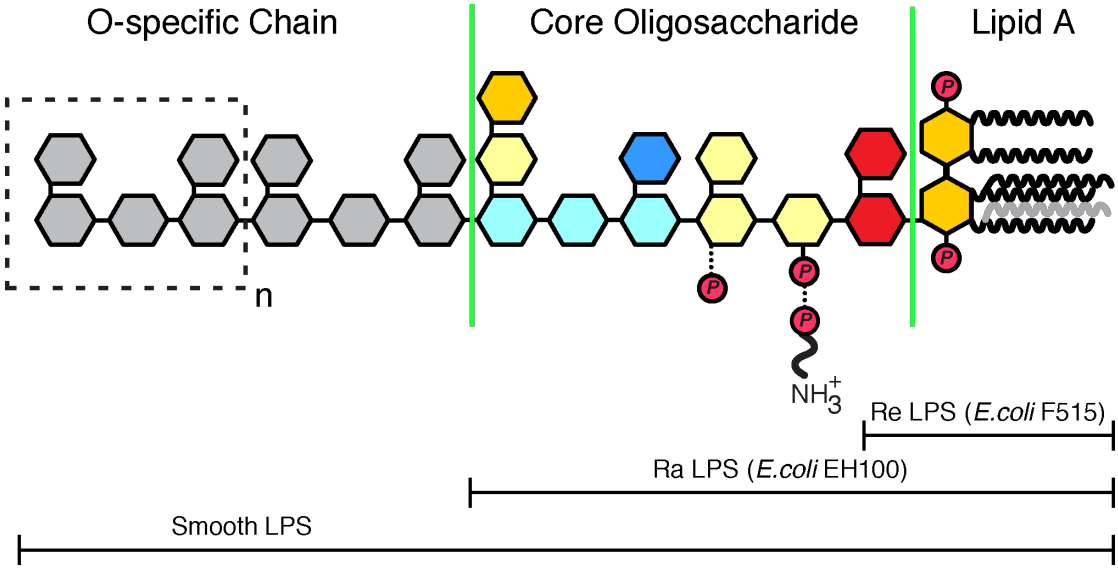

Figure 1: Architecture of *E. coli* LPS adapted from Wang et al. (14). Different sugar units are color coded as follows: dark yellow: glucosamine (GN); red: 3-deoxy-D-*manno*-oct-2-ulosonic acid (KD); light yellow: L,D-heptose; light blue: glucose; dark blue: galactose. Fatty acids attached to GN units are shown as wavy lines; red circles indicate phosphate groups. Dashed lines indicate variations in phosphorylation patterns, including a phosphoethanolamine. The lengths of LPS expressed by *E. coli* strains F515, EH100, and S-form (smooth-form) are shown as horizontal bars. LPS derived from F515 is predominantly pentaacylated. The respective missing fatty acid is shown in gray.

The immunostimulatory activity of LPS resides in the lipid A, which acts as a potent elicitor of the innate immune system by TLR4 or intracellular receptors (15). While this leads to a significant enhancement of the resistance to infection and is beneficial to the host, an uncontrolled and massive immune response caused by circulation of a large amount of LPS results in severe symptoms of sepsis, septic shock and multi-organ failure (16, 17).

Chemical variations of lipid A affects its capacity to interact and activate receptors of the immune system, and it has been shown that its intrinsic conformation is responsible for the agonistic and antagonistic activity (18, 19). The bis-phosphorylated lipid A backbone, with an asymmetric (4 + 2) distribution of six acyl chains, is the most potent agonist of the innate immunity in humans (20, 21); different variations from this chemotype are less or not agonistically active. The main elements that influence lipid A toxicity are the number and the distribution of acyl chains, the phosphorylation pattern and the presence of charged groups on the polar heads (22).

The structural elucidation of lipid A being of pivotal importance for the understanding of its biological properties, is hindered by its amphiphilic nature, namely the presence of hydrophilic and hydrophobic groups conferring a tendency to form micelles with low solubility in any solvent system. The current approach for the structural characterization of LPS makes use of chemical analysis, mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy (23). While MS allows determination of the differing numbers and nature of acyl residues and polar heads, and their distribution on the disaccharide backbone, NMR provides additional information on chemical nature, structure, and dynamics of LPS and its constituents (14).

Role of the LPS-binding protein (LBP) in LPS recognition

Cellular recognition of LPS involves the direct binding of its monomers to the TLR4-MD-2 receptor complex (24, 25), as illustrated in Figure 2 (26). LPS aggregates, however, have to be monomerized by LPS-binding protein (LBP) and cluster of differentiation 14 (CD14) (27).

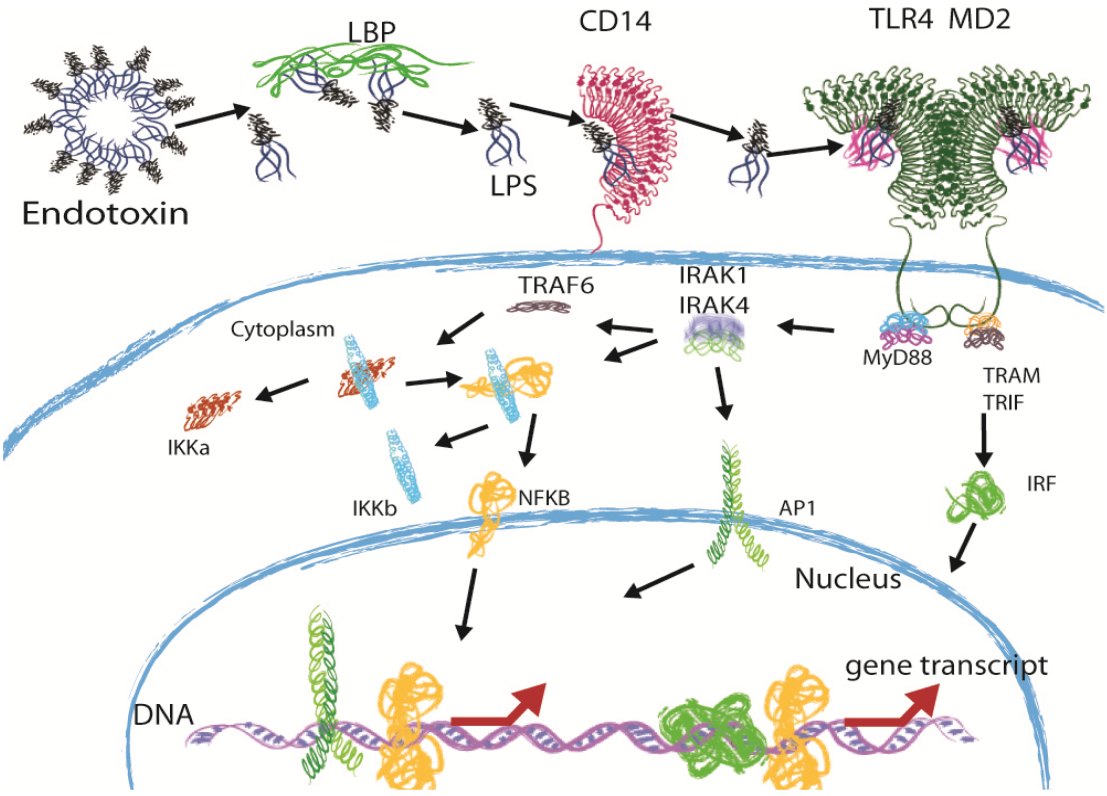

Figure 2: Cellular recognition of LPS adapted from Nijland et al. (26). LPS multimer is bound by LBP and its monomers are transferred to CD14. MD-2 binds to LPS and forms LPS-MD-2-TLR4 complexes. Dimerization of two of these complexes occurs, leading to the recruitment of adapter molecules: MyD88, the TIR-domain containing adapter protein-inducing IFN-β (TRIF) and the TRIF-related adapter molecule (TRAM). A signaling cascade is initiated, which leads to the release of the transcription factor NF-KB. NF-KB then moves into the nucleus and starts transcription of pro-inflammatory cytokines, such as IL-6 and TNF-α. *Via* a different pathway, initiated by different adapter molecules, type I interferon genes are also transcribed, leading to the production of IFN-α/β.

LBP is a ~60 kDa glycoprotein synthesized in hepatocytes and released into the bloodstream upon acute-phase inflammatory response (28, 29). It belongs to the lipid-binding/transfer protein family that includes bactericidal/permeability increasing protein (BPI), phospholipid ester transfer protein (PLTP), and cholesterol ester transfer protein (CETP) (30-32). Among these family members, BPI and LBP share the highest sequence identity (45 %) but markedly different functional roles. Whereas BPI neutralizes LPS and has potent bactericidal activity, LBP can transfer LPS to its cell-surface receptor and enhances cellular inflammatory responses to LPS (33, 34).

The first crystal structure of one of the family members, the human BPI (*h*BPI) (35), revealed two very similar, mixed α/β N- and C-terminal domains connected into a boomerang shape by a central β-sheet domain. The recent crystal structure of the mouse LBP (*m*LBP, Figure 3) (33), however, showed that while the individual domains are highly similar, the C-terminal domain of LBP is rotated relative to that of BPI. In addition, the *m*LBP crystal structure revealed a large groove near the C-terminal. These structural differences probably underlie the functional difference between LBP and BPI in mediating host defenses against Gram-negative bacteria (36).

A conserved cluster of positively charged residues (Arg^{119}, Lys^{120} and Lys^{124}, also shown as A patch in Figure 3A) in the N-terminal region of LBP has been identified by mutagenesis as part of the LPS-binding site (37). Recently, it has been shown that the positively charged residues from the A patch are necessary but insufficient to maintain high-affinity binding of LPS to LBP (38). Mutagenesis showed that additional positively charged residues (Lys^{67}, Lys^{69} and Arg^{173}, shown as B patch in Figure 3A) are essential for LPS binding. While the mutation of either the A or B patch retains partial binding, mutation of both patches completely abolishes it. The electrostatic surface of *m*LBP (Figure 3B) confirms the accumulation of positive charges at the N-terminal, which likely attract the negatively charged phosphates of LPS.

Using native polyacrylamide gel electrophoresis and sucrose density gradients, it was shown that LBP forms a high affinity complex with LPS and subsequently catalyzes the transfer of the monomers to a binding site on CD14 (39, 40). The transfer process exhibits first order kinetics, with even substoichiometric amounts of LBP being responsible for the movement of monomers from the aggregates to CD14 (41). Two models were proposed for explaining the LPS transfer to CD14 mediated by LBP: the "binary complex" model suggests that LBP first dissociates one LPS molecule from the micelle and, in a second bimolecular reaction, this complex binds to CD14 (40). The "ternary complex" model proposes a trimolecular complex formation involving LBP, LPS micelle and CD14 during the LPS monomer transfer to CD14.

Recently, negative-stain electron microscopy has been used to directly visualize the binding of LBP to the surface of LPS micelles via its N-terminal tip (38). Moreover, new motifs were also identified in the LBP C-terminal domain that are essential for LPS transfer. Briefly, complementary charge interactions between the C-terminal tip of LBP (Lys^{319} and Arg^{322}, also shown as C patch in Figure 3A) and a concave patch in CD14 C-terminal domain are required for the delivery of LPS from LBP to CD14. In addition, Asp^{311} in the LBP D patch near the C-terminal groove (Figure 3A) is necessary for the proper dissociation of CD14 from LBP-LPS complex after LPS transfer. Mutations of humans in this C-terminal region are linked to higher mortality rates resulting from bacterial infections (33), thereby supporting its functional significance.

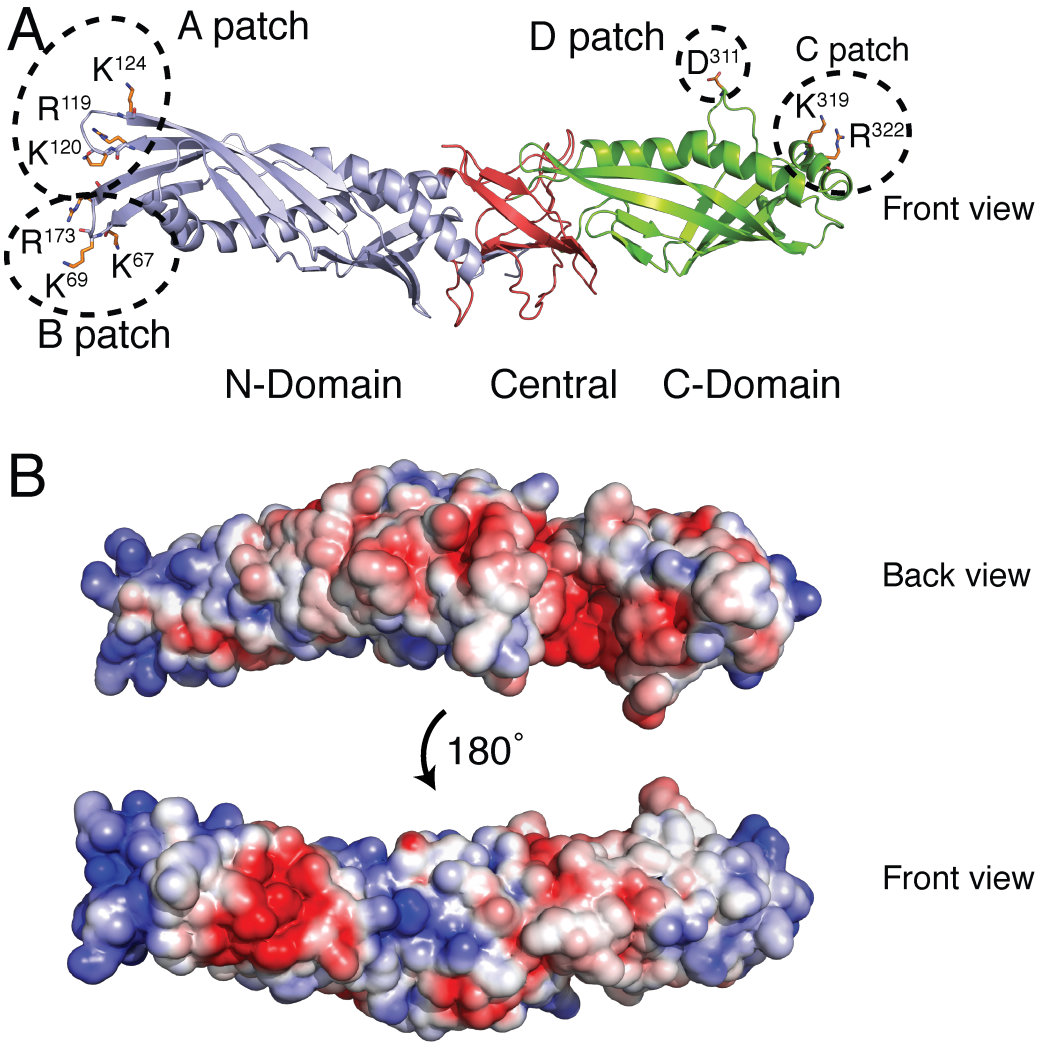

Figure 3: Crystal structure of the *m*LBP (PDB code 4M4D). **A** Basic residues in LBP's N-terminal domain (A and B patches) are important for LPS-binding; basic residues in the C-terminal domain (C patch) and acidic residue (D patch) are involved in LPS transfer to CD14. **B** Electrostatic potential representation of *m*LBP generated by the Pymol adaptative Poisson-Boltzmann solver (APBS). Blue: positive charges; red: negative charges; white: hydrophobic areas.

Although there has been tremendous progress in the structural characterization of the individual components required for the LPS transfer, including LBP, CD14, TLR4-MD2, and the TLR-MD2-LPS complex, the sequence and dynamics of the interactions of these components are not well understood. The transient nature of these interactions is advantageous for a rapid response of the immune system, but has made it difficult to elucidate the molecular structures of the transient intermediates (i.e., LPS-LBP, LPS-CD14, CD14-LBP-LPS, and CD14-LPS-TLR4-MD2 complexes). Here, we present data on the characterization of the human LBP (*h*LBP) produced using the baculovirus expression vector system (BEVS)/insect cells. We have characterized the recombinant protein using MS, size exclusion chromatography coupled to multi-angle light scattering (SEC-MALS), circular dichroism (CD) and NMR, and showed that the recombinant LBP is folded, monomeric, glycosylated, and competent for binding to LPS. However, the large size of LBP poses a significant challenge for NMR studies. We have developed a ^{15}N-labeling scheme for LBP, which produces an isotope-labeled sample with ~80 % labeling efficiency. The binding to LPS and precursors was studied by tryptophan fluorescence and SPR. The data suggest that LBP binds to aggregates of LPS with high affinity and coats their surface. This interaction has an electrostatic component, likely mediated by the positively charged residues on the N-terminal tip of LBP.

Results and Discussion

Cloning, expression, purification and biophysical characterization of hLBP

Construct design and small-scale expression optimization

Structural studies usually require milligram quantities of fully functional biomolecules. Heterologous expression of proteins with BEVS/insect cells provides not only good yields of the recombinant material, but also post-translational modifications (PTMs) very similar to those occurring in the native environment (42). The strategy used in this study, to express the recombinant *h*LBP, was based on two considerations. First, we wanted to use a system that had been successfully used to express other functionally active LPS-binding proteins (33, 38, 43-46). Second, we wanted to generate soluble *h*LBP as opposed to the inclusion body material produced when using *E. coli* in initial experiments (data not shown).

The construct for protein expression used in this study was modified from the transfer vector pACEBac 1 (Geneva Biotech) and was designed to express a tagged *h*LBP protein under the control of the polyhedrin promoter (Figure 4A). The signal sequence used for direct secretion of the recombinant protein into the culture media was from honeybee melittin (*Apis mellifera*, HBM) (47). The HBM signal sequence is cleaved upon secretion, leaving behind an Asp residue linked to the N-terminus of the mature *h*LBP. The C-terminus of the *h*LBP was linked *via* a Gly-Ser-Gly-Ser-Gly-Ser sequence to a ten-residue histidine tag (His_{10}) to facilitate purification by immobilized-metal affinity chromatography (IMAC). The Gateway-adapted BEVS (Invitrogen) transfer vector used in this study was designed to incorporate the target sequence *via* site-specific transposition into the genome of the bacmid. The final recombinant protein (from now on referred as LBP) expressed in insect cells consists of the sequence HBM-D-*h*LBP-GSGSGS-His_{10}.

The expression of LBP in insect cells was optimized in terms of viral titer and expression time using 25 mL test cultures. A viral titer of 10 mL/L gave consistently good results and was found to be the better compromise between virus amount and expression yield, with the optimal harvesting time being 60 hours post infection (Figure 4B). Subsequent Western blotting analysis (anti-human LBP antibody biG 42, Biometec) of the expression media revealed that LBP occurs as a mixture of two species migrating as ~50 and 55 kDa bands on a 4-20% polyacrylamide gel (SDS-PAGE). The higher molecular form is produced in much larger amounts and is a result of the glycosylation machinery of the insect cells, which recognizes the HBM secretory signal, and therefore produces a glycoprotein (discussed below).

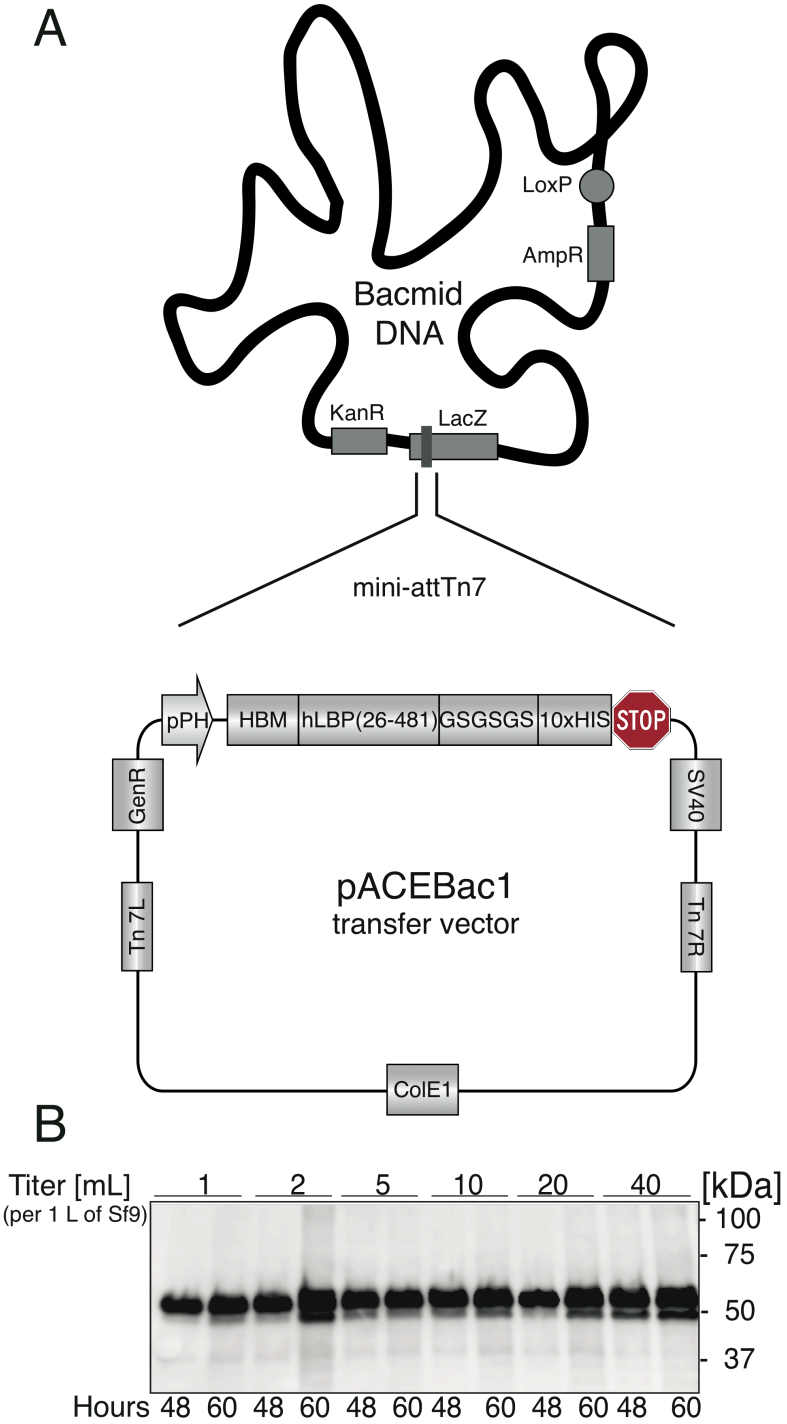

Figure 4: Cloning of the *h*LBP into the baculovirus expression vector. **A** Design of the expression construct showing the location of the HBM signal peptide, GS linker region and His_{10} purification tag upstream and downstream of the coding sequence for the LBP. **B** Small-scale cultures of Sf9 cells were infected with virus stock corresponding to the indicated functional titers of 1-40 mL per 1 L of cell culture. Expression levels of LBP were determined by Western blotting analysis (anti-human LBP antibody biG 42, Biometec) of the samples at the indicated time points.

Large-scale expression and purification

For the large-scale production of LBP, the size of the expression media was scaled to 1 L, and the relative viral titer was kept at 10 mL/L. The purification of the recombinant protein from insect cell medium was achieved in three chromatographic steps: first, ion exchange chromatography (IEX, Figure 5A) is employed to recover LBP from large media volumes, while simultaneously providing a first separation of the proteins based on their net charge. The elution profile upon application of the salt gradient shows that LBP is successfully eluted on an NaCl concentration range of ~0.2-0.5 M. The eluate from IEX is however not pure; it still contains a mixture of other higher and lower molecular weight proteinaceous contaminants. At this point, the His_{10} of the recombinant LBP allows the use of IMAC (Figure 5B), thus giving rise to a much purer protein sample eluted at an imidazole concentration range of 0.25-0.35 M. Finally, size-exclusion chromatography (SEC) (Figure 5C) is used to fully purify and evaluate the sample homogeneity. The SEC profile shows a homogeneous LBP eluted at ~15 mL of a Superdex 200 10/300 column (GE Healthcare), which corresponds to a molecular weight between ~35-67 kDa. The established purification procedure yields approximately 2-2.5 mg of highly pure and homogeneous LBP per liter of insect cell culture (Figure 5D).

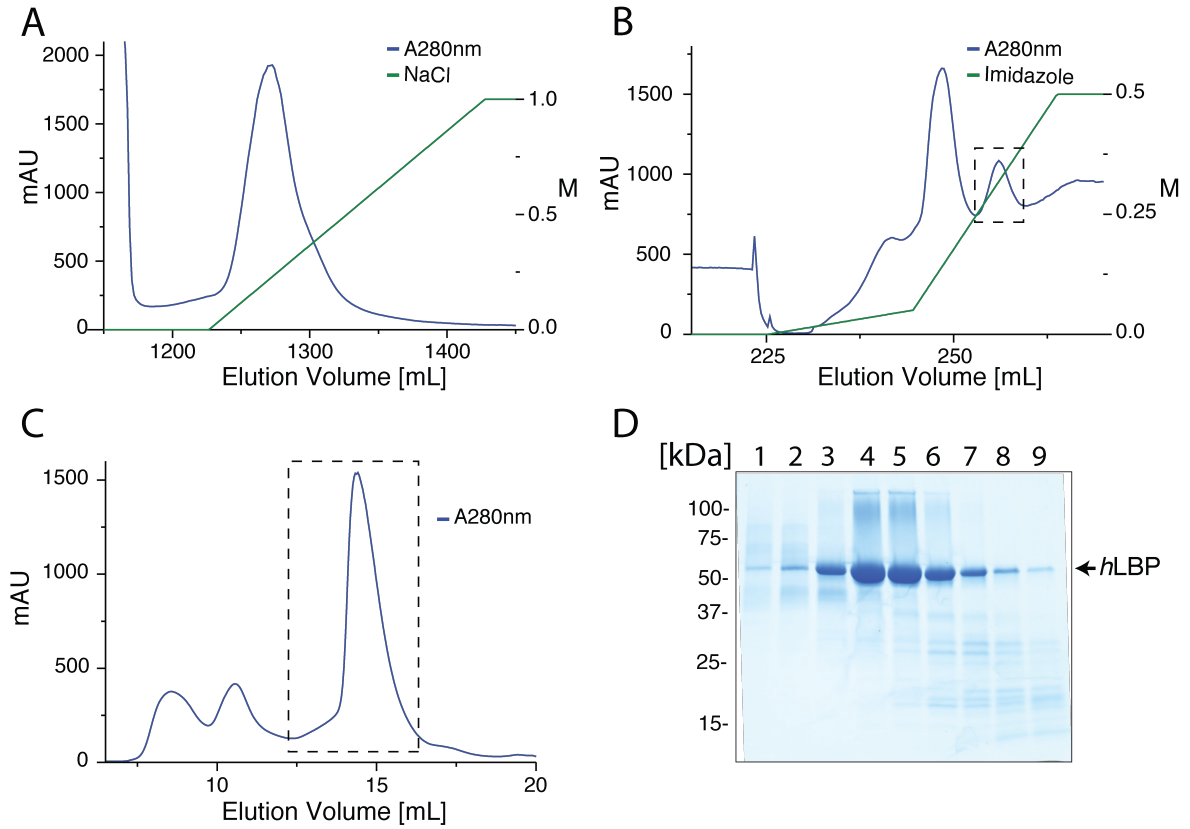

Figure 5: Large-scale purification of the LBP from 1 L insect cell medium, using a three-step chromatographic procedure. **A** IEX showing the profile of all proteins captured from the cell culture medium and eluted from the column using a linear gradient from 0 to 1 M NaCl. **B** IMAC profile of the pooled fractions resulting from the IEX step. A step-wise gradient from 0 to 0.05 M and from 0.05 to 0.5 M of imidazole was applied to the captured proteins, resulting in the separation of LBP (dashed box) from main impurities. **C** SEC profile of fractions collected in an imidazole concentration range of 0.25 to 0.35 M during the IMAC step. LBP is eluted in the region enclosed by the dashed box. **D** SDS-PAGE analysis of the fractions 1-9 collected between 12.5 and 17 mL of the SEC elution.

Analysis of LBP's post-translational modifications (PTMs)

In mammalian cells, LBP undergoes PTMs by addition of sugar groups to Asn^{300}, Asn^{355}, Asn^{386} and Asn^{394} (UniprotKB). In order to investigate the glycosylation state of LBP, the purified sample was subjected to PNGase F (NEB), which hydrolyzes complex-type N-linked oligosaccharides between the proximal N-acetylglucosamine (GlcNac) and the asparagine residue of the recombinant protein (Figure 6A). PNGase F treatment caused a downward shift of ~4-5 kDa as seen in Figure 6B, confirming that the LBP produced in insect cells is glycosylated.

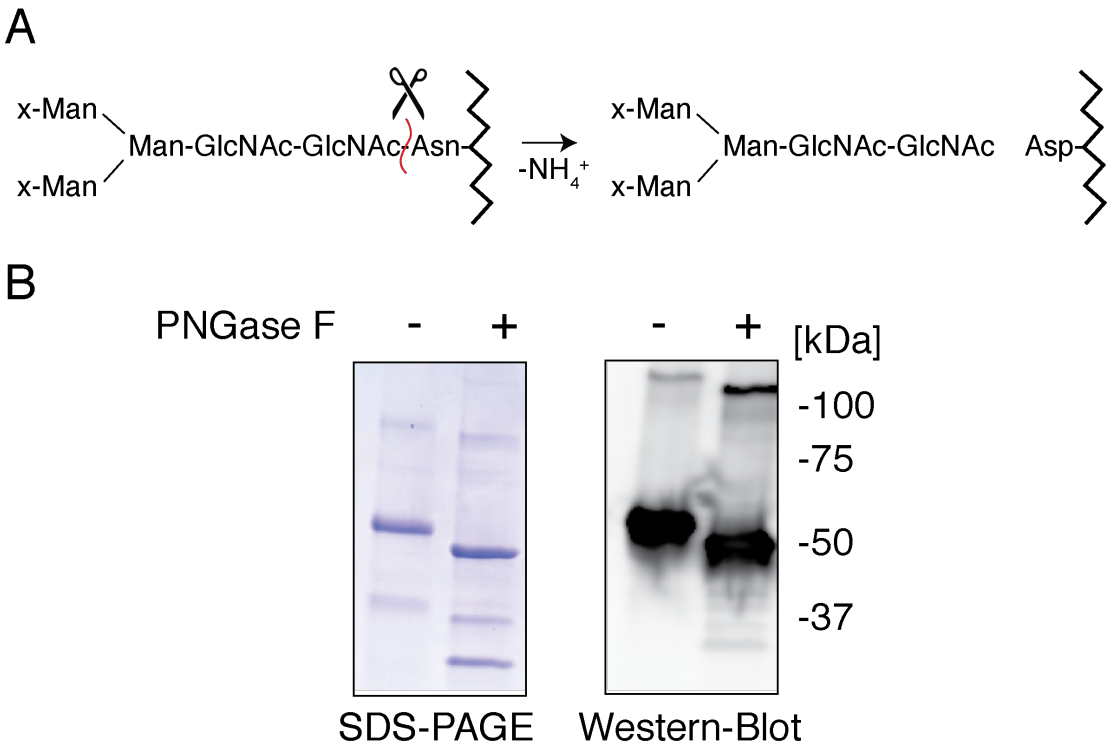

Figure 6: Verification of the glycosylation state of LBP. **A** Scheme of PNGase F cleaving between the N-acetylglucosamine and the Asn residue of the protein. **B** SDS-PAGE to verify the cleavage after incubation of an aliquot of 20 µL of baculovirus-infected Sf9 culture supernatant with 500 units of PNGase overnight at 25 °C. SDS-PAGE, stained by coomassie brilliant blue (CBB), and Western blot (anti-human LBP antibody biG 42, Biometec).

The follow-up MS analysis of the PNGase-treated/untreated forms of LBP allowed the identification of the sugar-modified sites (Table 1). The result confirmed three (Asn^{300}, Asn^{386} and Asn^{394}) out of the four glycosylation sites predicted in the UniProtKB entry P18428. The fourth (Asn^{355}), however, was not glycosylated according to MS. This finding is likely a consequence of the different N-glycosylation machineries of insect and mammalian cells (48).

Table 1: Identification of the sugar-modified sites in the LBP primary sequence by PNGase treatment and MS analysis. PNGase converts Asn to Asp at the site of glycosylation, MS is subsequently used to identify the positions.

Peptide	Modification (peptide)	Modification (*h*LBP)
LSVATnVSATLTFnTSKITGFLKPGK	N^{6} (N->D), N^{14} (N->D)	N^{386}, N^{394}
LARLYPNmnLELQGSVPSAPLLnFSPGN-LSVDPYmEIDAFVLLPSSSKEPVFR	N^{9} (N->D), N^{23} (N->D)	N^{336}, N^{350}
MVYFAISDYVFNTASLVYHEE-GYLNFSITDDmIPPDSnIR	N^{38} (N->D)	N^{313}
mVYFAISDYVFnTASLVYHEE-GYLnFSITDDMIPPDSNIR	N^{12} (N->D), N^{25} (N->D)	N^{387}, N^{300}

An accurate determination of the molecular weight of the glycosylated LBP could be obtained by MS (Figure 7A) and SEC-multi-angle light scattering (SEC-MALS) (Figure 7B). Both methods show an agreement on the size of the recombinant protein of about 57 kDa. Moreover, MS shows other species with similar molecular weights, which indicates that there might exist different levels of glycosylation in LBP. The most abundant species (56847.6 Da) possess ~5 kDa of total sugars, which is less than the ones from the LBP isolated from acute phase serum (~8 kDa) (28). Additionally, SEC-MALS shows the profile of a homogenous and monomeric protein sample. Overall, the data suggest that the lower molecular band observed in the supernatant of the insect medium from the small-scale expression optimization in Figure 4B consists of non-glycosylated or partially-glycosylated form of LBP, while the higher molecular weight band (Figure 4B and 5D) consists of higher-glycosylated forms.

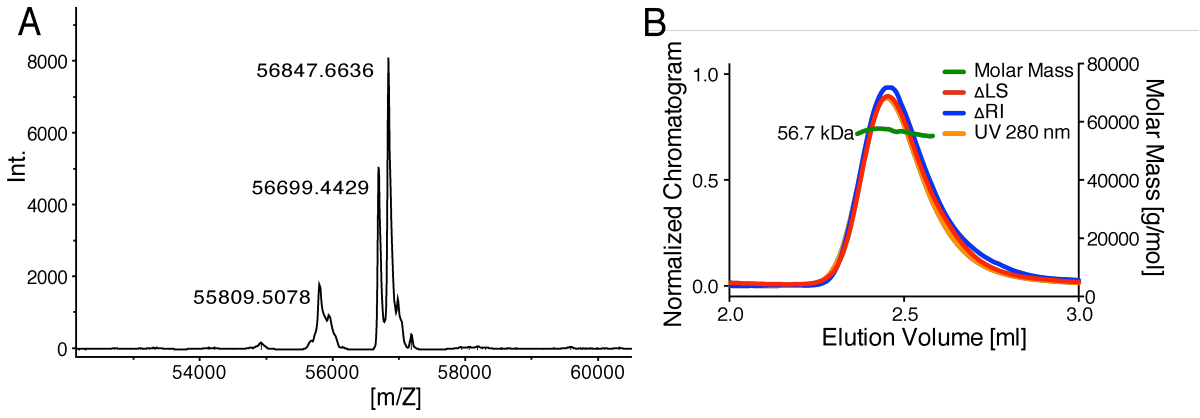

Figure 7: Characterization of the size and homogeneity of the insect cell-produced LBP. **A** MS of 10 µM LBP shows different levels of glycosylation. Buffer: 5 mM HEPES, 200 mM NaCl, 1 mM EDTA, pH 7.5. **B** SEC-MALS of 16 µM LBP in 5 mM HEPES, 500 mM NaCl, 1 mM EDTA, pH 7.5 confirms the molecular weight of the most abundant LBP species and shows a monomeric elution profile. ΔLS is the light scattering detector; ΔRI is the refractive index detector.

Secondary structure analysis

CD spectroscopic analysis was carried out on LBP to confirm the nature of its structural fold. The far-UV CD spectrum shows a minimum at 216 nm and a maximum at ~195 nm (Figure 8), features characteristic of a folded β-sheet-like protein. Subsequent fitting of the data using Dichroweb (49, 50) revealed that LBP contains a mixture of α-helix and β-sheet.

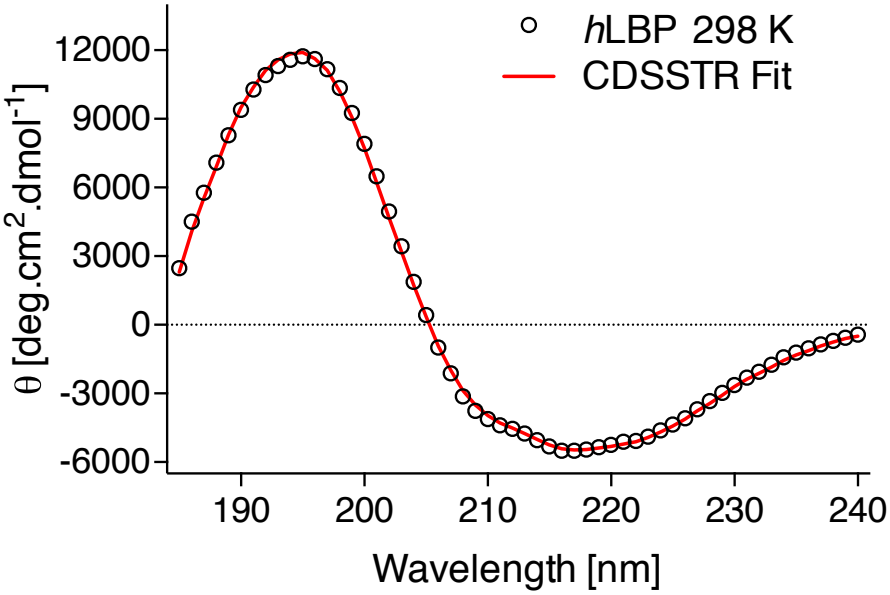

Figure 8: CD spectrum of 5 µM LBP sample recorded at 25 °C and pH 7.5. Experimental data was fitted online (49) with the CDSSTR algorithm.

A comparison of the secondary structure composition derived from CD and the *m*LBP crystal structure (33) is shown in Table 2. Both data agree on the β-strand content (~40 %) and the amount of disorder (~30 %). Small discrepancies are observed for the helical content, which is 17 % for *m*LBP and 6 % for *h*LBP, and also for the turns, which is 7 % for *m*LBP and 21 % for *h*LBP. These data indicate that both proteins share a similar secondary structure, which is expected from their high sequence identity of 69.2 %.

Table 2: Secondary structure composition of LBP, as determined from CD and the crystal structure (PDB code 4M4D).

	*m*LBP[a]	*h*LBP[a]
% Helix	17	6
% Strand	42	41
% Turn	7	21
% Unordered	34	31

[a] Percentage of residues in secondary structure elements found in the crystal structure (33). The values were derived by analysis of the PDB structure 4M4D using the DSSP software (51).
[b] Percentage of residues in secondary structure elements, as derived from the fit of the experimentally measured CD data.

The protein folding status was further assessed by NMR using a 1-1 echo sequence (52). The one-dimensional ^{1}H NMR was analyzed for signal dispersion in the indole, aromatic, and amide region (6-11 ppm), and suggested presence of protein secondary structure (Figure 9). The broad lines observed are caused by the large size of LBP. The amide $^1H^N$ T_2 was estimated by recording the 1-1 echo sequence at two different relaxation delays (0.1 and 2.9 ms), and yielded 5.1 ms; this small value of T_2 imposes a challenge for recording high-dimensional spectra of LBP.

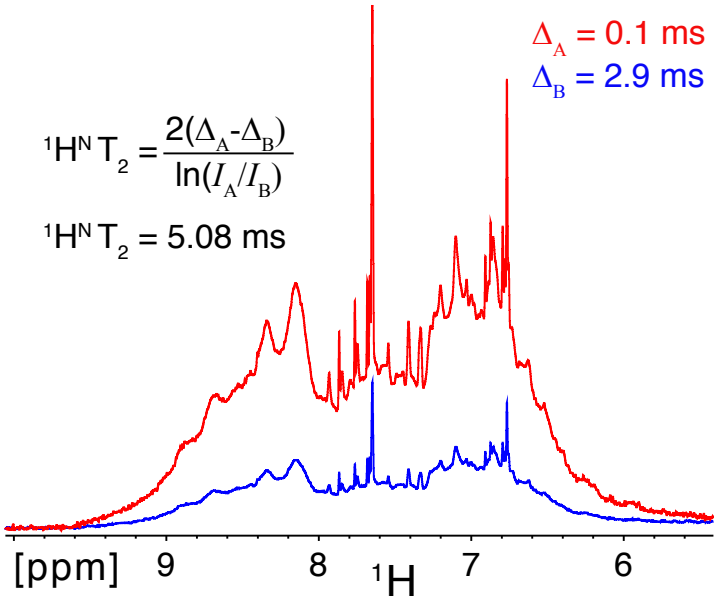

Figure 9: The 1-1 echo sequence of 100 µM LBP recorded at 37 ºC. The spectral dispersion of the amide resonances suggests that LBP is folded. The 1-1 echo sequence recorded at 0.1 and 2.9 ms relaxation times allowed estimation of the $^1H^N$ T_2 time.

Further confirmation of the folded state of LBP was obtained from the ^{1}H-^{1}H NOESY spectrum (Figure 10), where cross peaks arise from dipolar couplings between spins *via space* (<5 Å). Despite the inability to resolve the individual cross peaks due to low resolution, the NOE connectivities in the H^N-H^N region indicate the presence of formed α-helices. The presence of β-sheets is usually recognized by strong H^α-H^α NOEs. However, this region is obscured in the NOESY spectrum due to the 1-1 sine modulated detection mode of this experiment. Therefore, signals between 4.5 and 5.5 ppm are rendered invisible. Finally, the upfield-shifted methyl resonances appearing below 0 ppm also indicate the folded state of the protein.

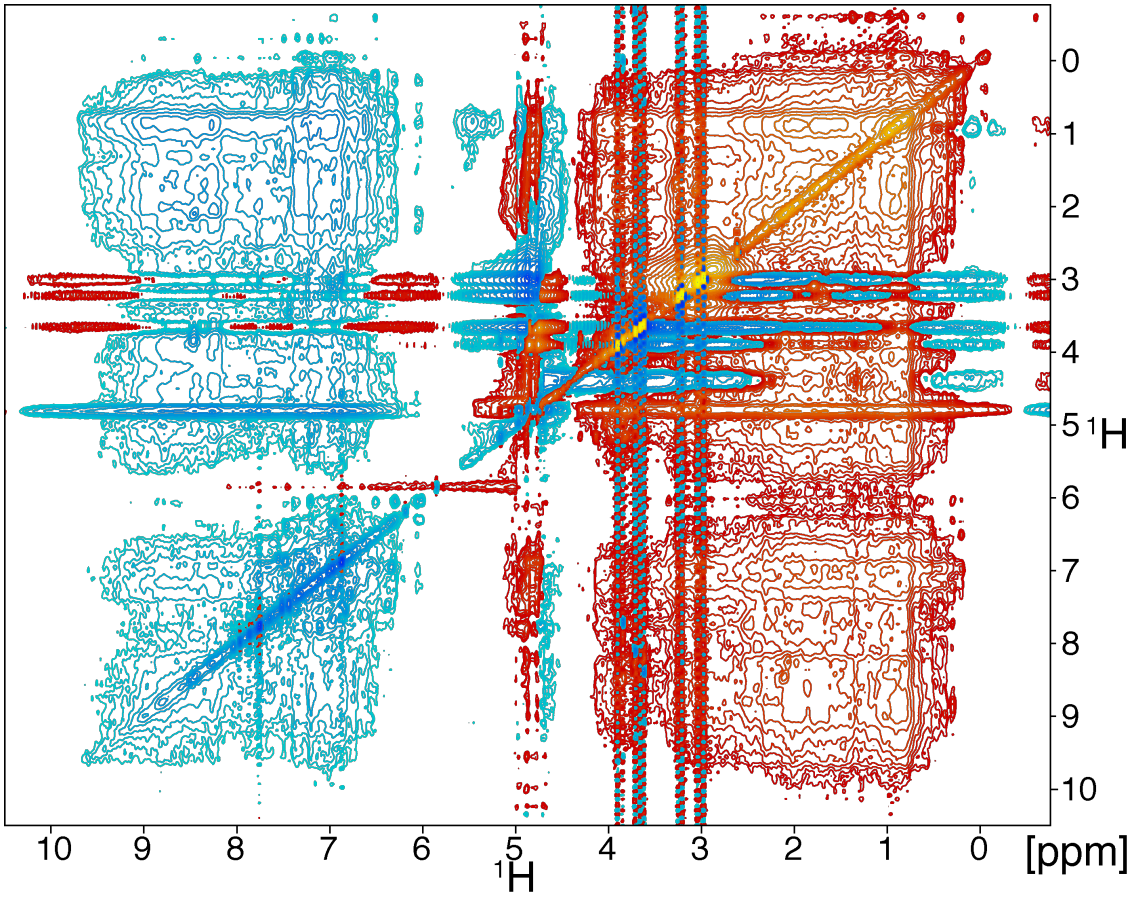

Figure 10: 2D-NOESY, with 1-1 detection, spectrum of 100 µM LBP sample recorded at 37 ºC, for 9 hours and 36 minutes on a 900 MHz spectrometer.

^{15}N-isotope labeling of LBP in insect cells

A labeling strategy based on a previously reported recipe (53) was developed for the expression of LBP in insect cells using isotope-labeled yeast extract (YE). The complete procedure is described in the PhD thesis of Christian Opitz (54). Briefly, the production of *Pichia pastoris* YE is achieved as follows: first, cells are grown in a fermenter under fed-batch conditions with typical volumes of 0.5-5 L. Growth on a glucose- and ^{15}N ammonium-based minimal medium yields up to 64 g cell wet weight of yeast biomass per liter of cell culture. Subsequently, an optimized protocol for *Pichia pastoris* autolysis was developed, which maximizes the amino acid content in the YE. The protocol yields on average 6.6 ± 0.7 g lyophilized YE per liter of cell culture, containing 38 ± 4 % of free amino acids. This material is supplemented to insect cell culture medium in order to produce labeled proteins.

Sf9 insect cells maintained in SF-4 (Bioconcept) medium were either kept in SF-4 or exchanged to Insect Xpress (Lonza) commercial media to establish LBP reference expression. While an average of 2.4 mg/L of purified LBP was obtained in SF-4 (Figure 11, condition 1), expression in Lonza resulted in 50 % lower yield (Figure 11, condition 2). This decrease in yield is likely a consequence of the different medium composition. Whereas SF-4 is solely based on YE, Insect Xpress contains an additional hydrolysate of undisclosed composition (personal communication with supplier's scientific report). Therefore, we decided to use SF-4-adapted cells for labeling in amino acid-depleted SF-4 based medium. Hence all the following labeling schemes were performed with SF-4-adapted cells.

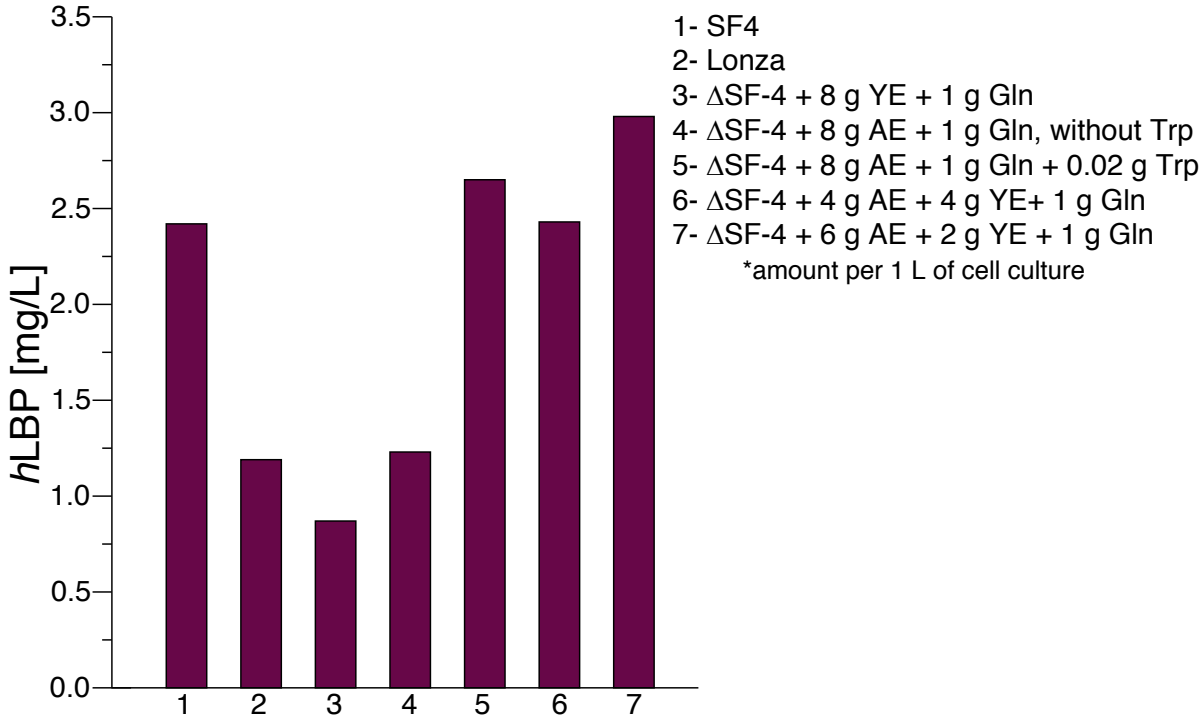

Figure 11: Expression yield of the *h*LBP produced in Sf9 insect cells under various conditions. Reference expression was established in commercial SF-4 medium (1). Additionally, expression in commercial Insect Xpress medium (2) following adaptation to SF-4 medium was carried out. Different supplements to SF-4 medium depleted in amino acids and yeastolate (ΔSF4) were tested (3-7).

For isotope labeling, amino acid-depleted, yeast extract-free SF-4 medium (ΔSF4) was supplemented with 1 g/L of glutamine and 8 g/L of YE and the expression of LBP was analyzed (Figure 11, condition 3). Surprisingly, only 0.9 mg/L of purified material was obtained. This significant reduction had not been observed for the previously expressed proteins.

To prevent the drop in yield, a commercial algal extract (AE, ISOGRO, Sigma) was used as source of amino acids in ΔSF4, as recently reported (55). Expression of LBP in ΔSF4 supplemented with 8 g/L of AE and 1 g/L glutamine resulted in a yield of 1.2 mg/L of protein (Figure 11, condition 4).

Restoration of the expression yield to 2.7 mg/L of purified LBP was obtained by supplementing 1 g/L glutamine and 20 mg/L of tryptophan to the AE-containing expression medium (Figure 11, condition 5). The decision was based on the fact that the amino acids cysteine and tryptophan are lost during conventional acid hydrolysis for the preparation of commercial algal extract (AE), while glutamine and asparagine are converted to glutamic acid and aspartic acid, respectively (56).

To optimize the labeling, a strategy based on the combination of YE and AE was tested. This combinatorial approach might compensate the lack of tryptophan in the commercial AE and, perhaps, cure the negative impact of YE which may be due to an inhibitory factor influencing the production or secretion of LBP. Therefore, ΔSF4 was supplemented with 1 g/L of glutamine, without addition of tryptophan, and two different ratios of AE/YE accounting for a total extract concentration of 8 g/L. Supplementing ΔSF4 with 4 g/L of AE and 4 g/L of YE yielded 2.4 mg/L of purified LBP (Figure 11, condition 6). On the other hand, blending 6 g/L of AE with 2 g/L of YE yielded 3.0 mg/L of recombinant protein (Figure 11, condition 7).

From the limited number of experiments, the following conclusions can be drawn:

(1) the low expression yields observed for the commercial AE can be recovered to levels similar to those observed in SF4 (Figure 11, condition 1) by supplementing tryptophan and glutamine to the growth medium;

(2) the combination of AE, YE and glutamine produces an outcome similar to supplementation of AE with tryptophan and glutamine;

(3) the data here presented is insufficient to draw extensive conclusions about the inhibitory effects of YE. However, if existent, this inhibition is not effective at concentrations until 2 g/L YE, as the resulting yield is similar to that obtained in SF4 medium (Figure 11, condition 1);

(4) a more thorough analysis of the effects of increasing the amount of AE or decreasing the amount of YE needs to be carried out in order to better understand the low yields observed for YE alone.

Large-scale production and NMR characterization of the uniformly ^{15}N-labeled LBP

For the NMR analysis, ^{15}N-labeled LBP was expressed in Sf9 insect cells using a 1:1 ratio of ^{15}N-labeled AE/YE as described above and reported by Opitz et al. (54). For this, ΔSF4 was supplemented with 1 g/L of unlabeled glutamine, 4 g/L of ^{15}N-labeled AE (N-ISOGRO, Sigma) and 4 g/L of ^{15}N-labeled YE. After purification, 2.7 mg/L of ^{15}N-labeled LBP was obtained. Since it was previously observed that LBP concentrations exceeding 8 mg/ml (~150 µM) led to irreversible precipitation of the protein, only a fraction of 1.5 mg was used to prepare the 270 µL NMR sample. Figure 12 shows the resulting ^{1}H-^{15}N TROSY HSQC recorded for 60 h on the 111 µM ^{15}N-labeled LBP. In total, 324 of the 473 expected backbone amide resonances were observable. To further characterize this sample, ^{15}N-incorporation of LBP was analyzed by a $^{1}J_{HN}$ 1D spin echo difference experiment yielding an estimated ^{15}N-incorporation of 79 %. As expected for a protein of ~57 kDa, including glycosylation sites, further improvement of the spectral quality would require deuteration.

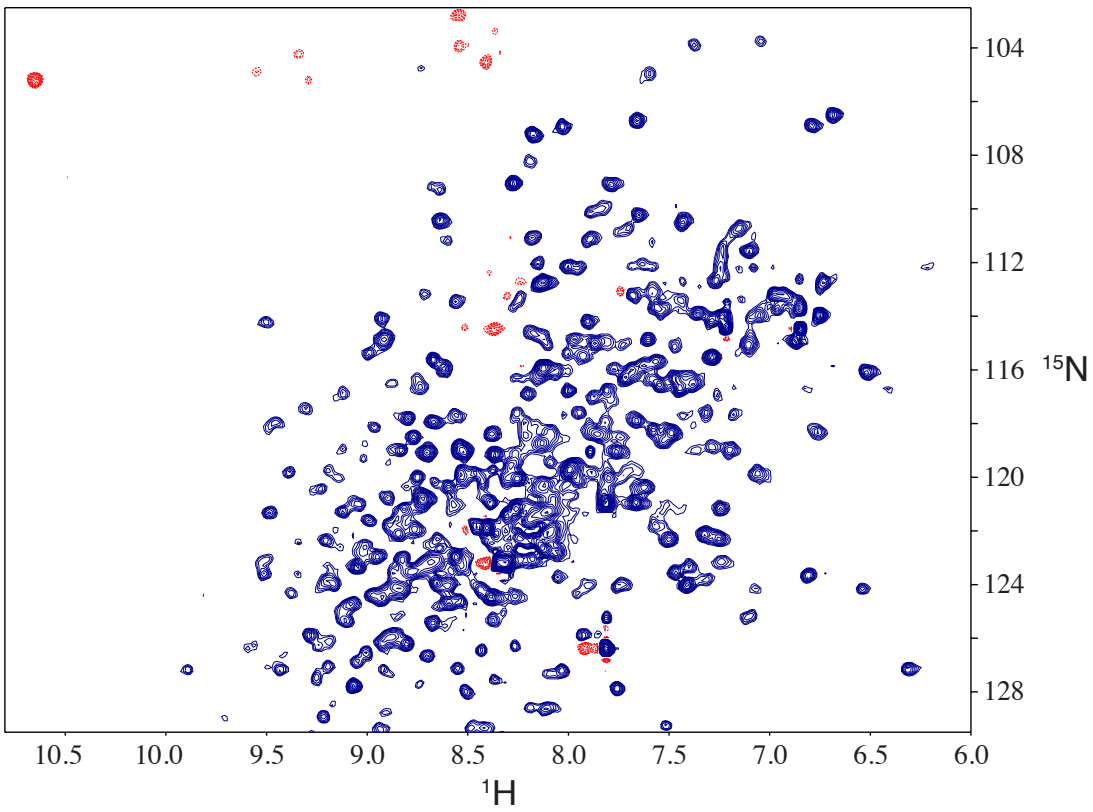

Figure 12: ^{1}H, ^{15}N-TROSY HSQC spectrum of ^{15}N-labeled LBP. ^{15}N-labeled LBP was expressed in Sf9 cells grown in a growth medium containing a combination of ^{15}N-yeast extract and ^{15}N-algal extract supplemented with unlabeled glutamine as described in the text. The spectrum was recorded in 60 h on a 111 µM sample of ^{15}N-labeled LBP on a 900 MHz NMR instrument equipped with a cryogenic probe.

To obtain information of the oligomeric state of LBP, ^{15}N T_1 and T_2 relaxation times of 2808 ms and 25 ms were measured using 1D ^{1}H-detected ^{15}N relaxation experiments (Figure 13) (57). Using the obtained values, the rotational correlation time (τ_c) could be derived by Eq. 1 (58):

$$\tau_c \approx \frac{\left(\sqrt{\frac{6T_1}{T_2}-7}\right)}{4\pi\nu_N} \qquad \text{(Eq. 1)}$$

where ν_N is the Larmor frequency of the ^{15}N nuclei.

We obtained a τ_c of 22.8 ns for LBP at 37 ºC, which agrees with the expected rotational correlation time of proteins of approximately 50 kDa. This indicates that LBP exists in solution as a monomer.

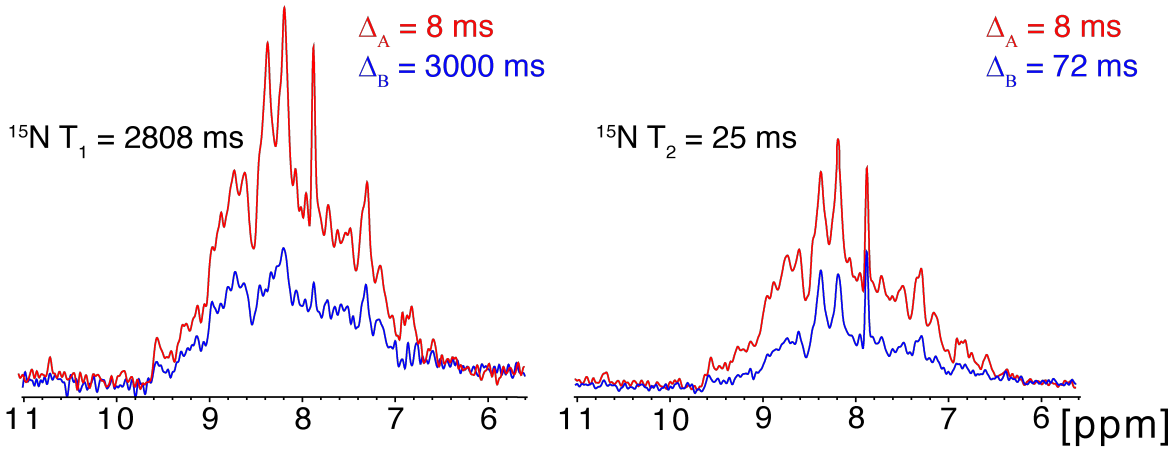

Figure 13: The determination of LBP relaxation times from 1D traces of the ^{15}N T_1 and T_2 relaxation experiments recorded using short (red) and long (blue) relaxation delays Δ at 37 ºC.

Local probing of the LBP-LPS interaction by the intrinsic tryptophan fluorescence

Characterization of the intrinsic tryptophan fluorescence of LBP

LBP contains two conserved tryptophan residues (Trp^{116} and Trp^{176}) embedded in its N-terminal tip (Figure 14A,B). After excitation at 295 nm, these two amino acids show a broad intrinsic fluorescence emission spectra with a maximum centered at ~320 nm (Figure 14C black), after subtraction of the Raman signal (59) from a separate buffer experiment. The LBP fluorescence spectrum is blue-shifted relative to free tryptophan, for which the emission maximum occurs at ~350 nm in aqueous solvent (Figure 14C blue). A red shift of the emission spectrum is well known as a solvent relaxation effect, where reorientation of the solvent molecules occurs around the fluorophore's excited-state dipole thereby lowering its energy (59). Therefore, one can conclude from the blue-shifted emission of the tryptophans in LBP that they reside in a more hydrophobic environment.

A comparison of the emission spectrum of 0.333 μM LBP (corresponding to 0.333 μM of Trp^{116} + 0.333 μM of Trp^{176}) with 0.333 μM of free tryptophan shows that the summed fluorescence intensity of LBP's tryptophans is quenched by a factor of four. A possible cause for the observed quenching of LBP fluorescence is the proximity of Trp^{116} and Trp^{176} to internal quenchers (59), such as the side-chain of His^{70} at a distance of 5.4 Å from Trp^{176}; the side-chain of Tyr^{76} at a distance of 13.5 and 10.6 Å from Trp^{176} and Trp^{116}, respectively; and the proximity between the tryptophans themselves, since they are at an average distance of 8.6 Å from each other.

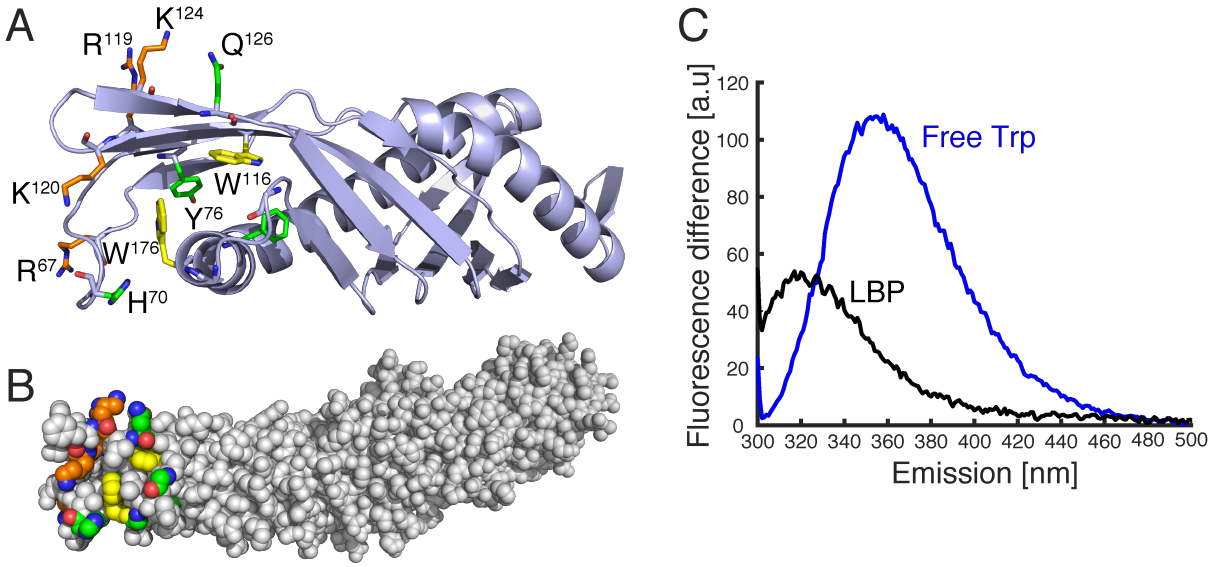

Figure 14: Usage of the LBP tryptophan fluorescence as a probe for the local environment. **A** Cartoon representation of the N-terminal domain of the *h*LBP modeled (SWISS-MODEL) using the *m*LBP crystal structure (PDB 4M4D). Tryptophans (yellow), basic residues (orange) and possible fluorescence quenchers (green) are shown as sticks. **B** Space-filling representation of the N-terminal domain of the *h*LBP model. **C** Emission spectrum of 0.333 µM LBP and 0.333 µM free Trp solution at 23 ºC after buffer subtraction. The excitation wavelength used was 295 nm, and the bandwidths of excitation and emission were 5 and 3 nm, respectively. The photomultiplier voltage was 550 V, and the scanning speed 200 nm/min at a step size of 1 nm. Both samples contain 5 mM HEPES, 500 mM NaCl, 1 mM EDTA, pH 7.5.

Tryptophan fluorescence changes upon LBP binding to LPS aggregates

Upon addition of increasing amounts of LPS F515 (chemical structure shown in Figure 16) to a 0.5 μM solution of LBP, a significant increase of its intrinsic fluorescence emission maximum was observed until a plateau was reached at 4.6 μM (Figure 15A). The very pronounced emission peak in Figure 15A with maximum centered at approximately 327 nm is a result of Raman scattering of the incoming light by the solvent. It is removed when subtracting the fluorescence spectrum of the buffer alone (Figure 15B). Apparently, the increase in fluorescence emission upon the addition of LPS is not accompanied by a shift of the emission wavelength (Figure 15B,D). The increase in fluorescence was quantified by integrals over the emission spectrum (Figure 15C), and the plot of the resulting values versus the glycolipid concentrations corresponds to a sigmoidal binding isotherm with a midpoint at about 2.5 μM. Clearly, the binding of LPS F515 reduces the quenching of LBP's tryptophan fluorescence by about 10 % relative to free tryptophan, which indicates local conformational changes at the N-terminal tip of LBP in the vicinity of the putative binding site comprising the cluster of positively charged residues (37, 38, 60, 61). On the other hand, the absence of changes of the maximum wavelength indicates that the polarity of the tryptophans's environment is not affected. Therefore, it is unlikely that Trp^{116} and Trp^{176} get further exposed to solvent upon LPS binding.

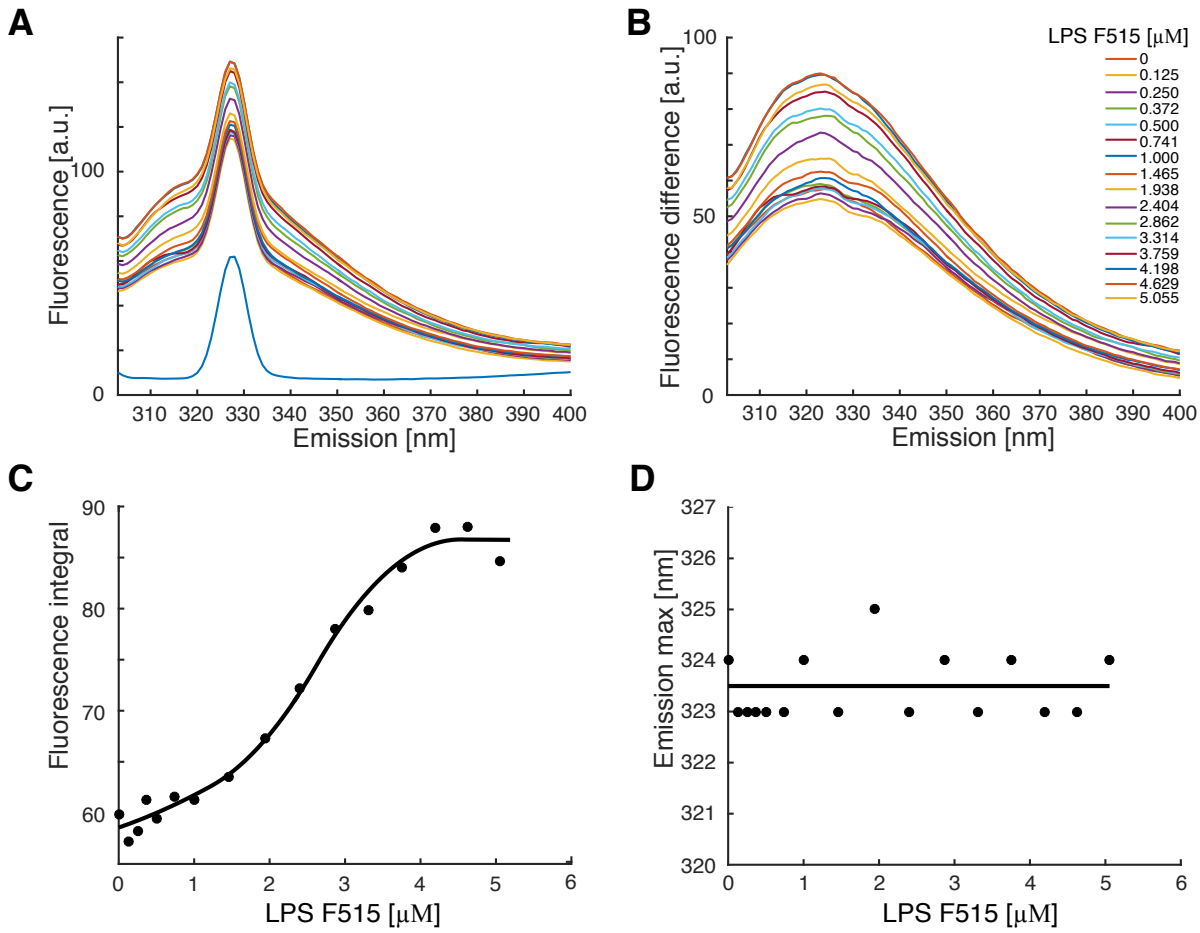

Figure 15: Effect of increasing concentrations of LPS F515 on the environment of LBP's tryptophan residues. **A** Emission spectra of a solution of 0.5 μM LBP upon addition of increasing concentrations (0, 0.125, 0.25, 0.372, 0.5, 0.741, 1, 1.465, 1.938, 2.404, 2.862, 3.314, 3.759, 4.198, 4.629 and 5.055 μM) of LPS F515, at 23 ºC. The blue line at the bottom is the (signal) spectrum of the buffer alone (see Figure 14), showing the Raman peak. **B** Emission spectrum after buffer subtraction. **C** Titration curve of the fluorescence integral between 308-318 nm and 335-380 nm. **D** Wavelength of the maximum emission as a function of the glycolipid concentration. All the details of the fluorescence measurements were as described in Figure 14.

In order to define the relative contribution of the lipidic or carbohydrate domains in the interactions with LBP, we analyzed the spectral properties upon interaction with various other structurally related glycolipids (chemical structures shown in Figure 16).

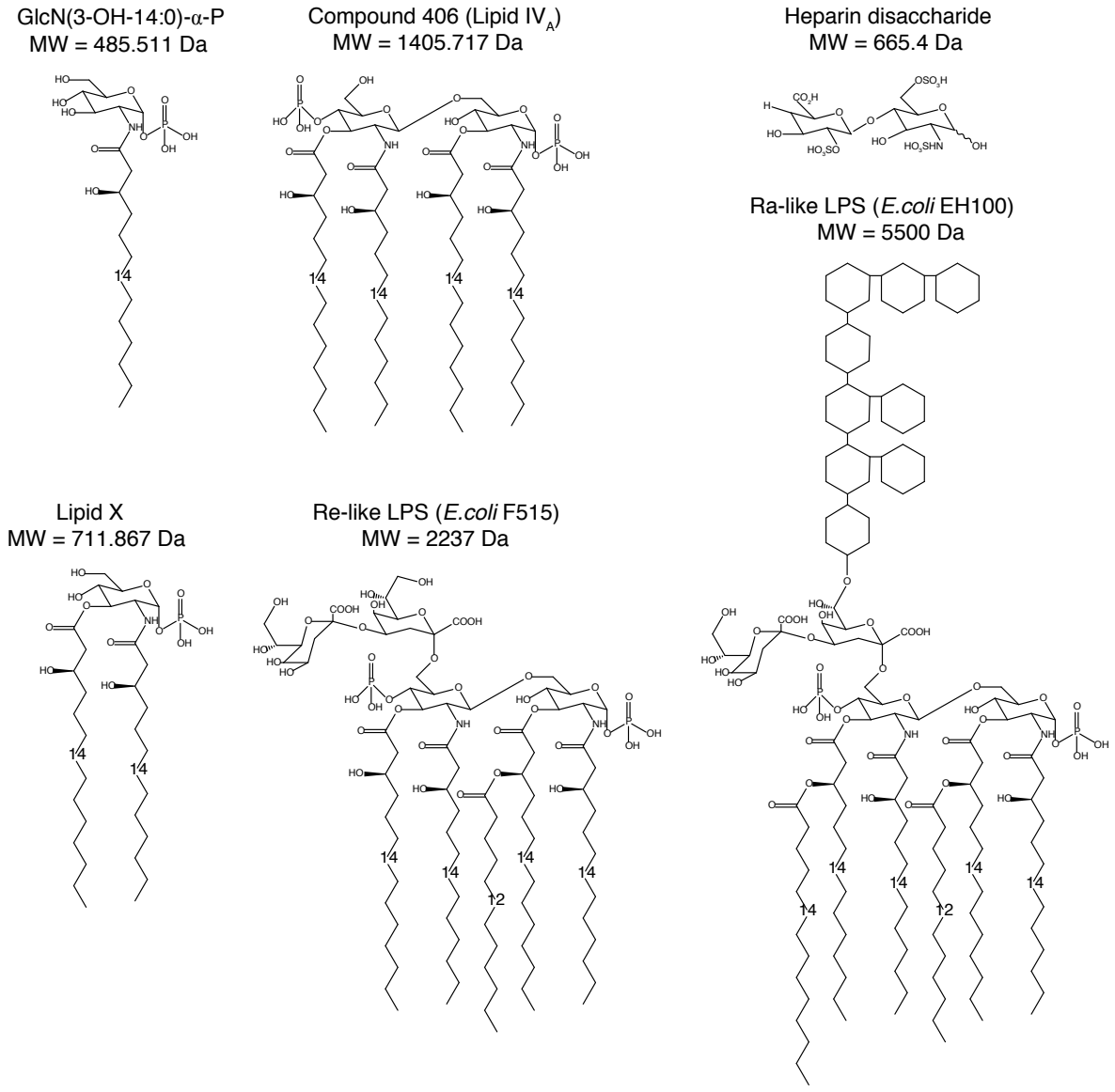

Figure 16: Chemical structures of the glycolipids used for the interaction studies with LBP. The glycolipids GlcN(3-OH-14:0)-α-P, Lipid IVa, Lipid X and Re-like LPS (*E.coli* F515) were kindly provided by Professor Ulrich Zähringer. The compounds Heparin disaccharide and Ra-like LPS (*E.coli* EH100), with product codes H1001 and L9641, were purchased from Sigma-Aldrich.

The result of the interaction of the LPS precursor Lipid X, the synthetic Lipid IVa, and the charged heparin disaccharide with LBP is shown in Figure 17 A,B. All the glycolipids promoted a significant increase of the fluorescence emission, the largest change was observed for LPS F515 and Lipid IVa. However, while the emission wavelength did not change upon titration of LPS F515 and Lipid IVa, Lipid X caused a red shift of 7.5 nm even at a concentration of 0.125 µM. Since this effect occurred well below the concentration of LBP (0.5 µM) and was concentration-independent, it is likely due to a contaminant in the Lipid X preparation.

Both heparin disaccharide and buffer alone (used as negative control) failed to produce any changes upon titration, except for a decay of the fluorescence signal, which was also observed even for free tryptophan in the absence of any glycolipids. This reduction is apparently due to photobleaching, since decreasing the excitation bandwidth reduced the effect, albeit at the expense of sensitivity.

The effect observed upon adding increasing amounts of the different compounds suggests that the increase in fluorescence emission is a direct consequence of the binding. The larger signal increase observed for LPS F515 and Lipid IVa compared to Lipid X is possibly a consequence of their larger size.

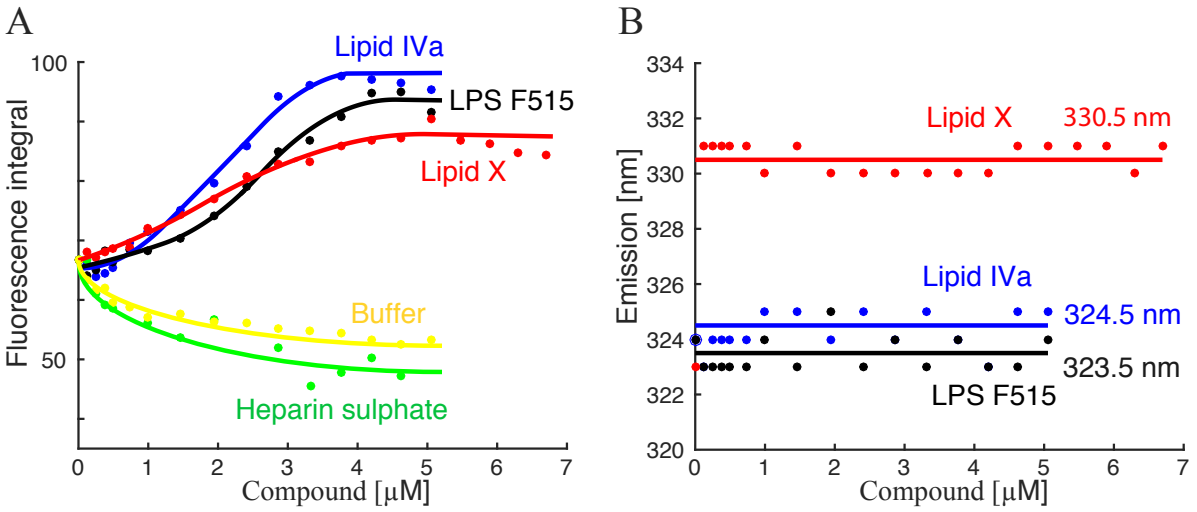

Figure 17: The effect of increasing concentrations of different glycolipid moieties on the fluorescence signal. **A** Plot of the integrated fluorescence signal (308-318 nm and 335-380 nm) of a 0.5 µM LBP solution, upon titration of increasing concentrations of the different ligands. **B** Plot of the wavelength of the maximum emission as a function of the glycolipid concentration. All the details of the fluorescence measurements were as described in Figure 14.

Investigation of the influence of the electrostatic interaction on the binding

To better understand the effect of the electrostatics on LBP-Lipid X binding, fluorescence measurements were also carried out in buffers with increasing concentrations of NaCl as shown in Figure 18. The binding-induced fluorescence response decreases continuously from 0.2 to 2 M NaCl, indicating the importance of electrostatics in the interaction.

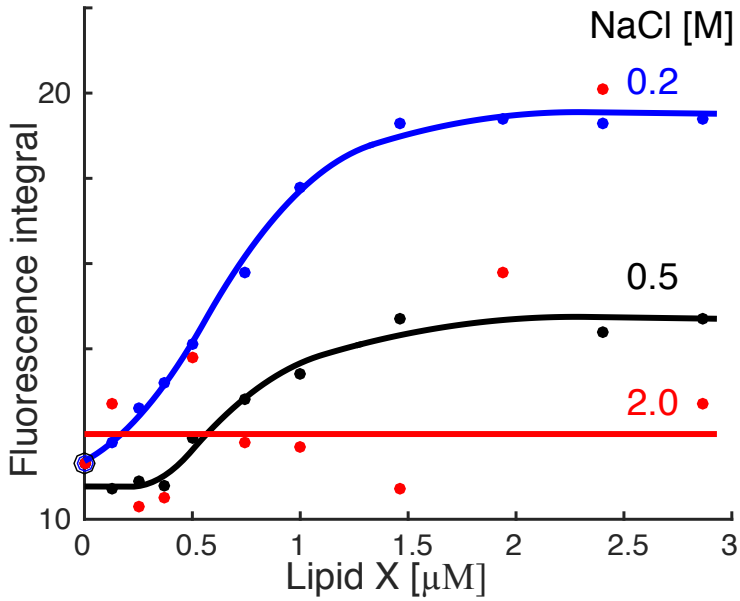

Figure 18: Effect of increasing concentrations of NaCl on the fluorescence signal. Plot of the integrated fluorescence signal (308-318 nm and 335-380 nm) of a 0.25 μM LBP solution at increasing concentrations (0, 0.125, 0.25, 0.372, 0.5, 0.741, 1, 1.465, 1.938, 2.404 and 2.862 μM) of Lipid X. All the details of the fluorescence measurements were as described in Figure 14.

Influence of mechanical stress on LBP probed by Trp fluorescence and CD

We observed that the fluorescence emission of LBP in solution increased continuously over the period of about an hour upon stirring the solution in the quartz cuvette with a magnetic stirrer (Figure 19A-C). The increase in florescence does not occur in the absence of stirring and could be explained by a reduction of quenching due to an increased distance of quenchers to the LPB tryptophan residues (Figure 14B). In addition, CD spectroscopy showed that the fluorescence increase is accompanied by a decrease of the molar ellipticity (Figure 19D), corresponding to a loss of secondary structure. Therefore, the data are consistent with unfolding upon stirring and also indicate that quenching is an intrinsic property of the folded LBP.

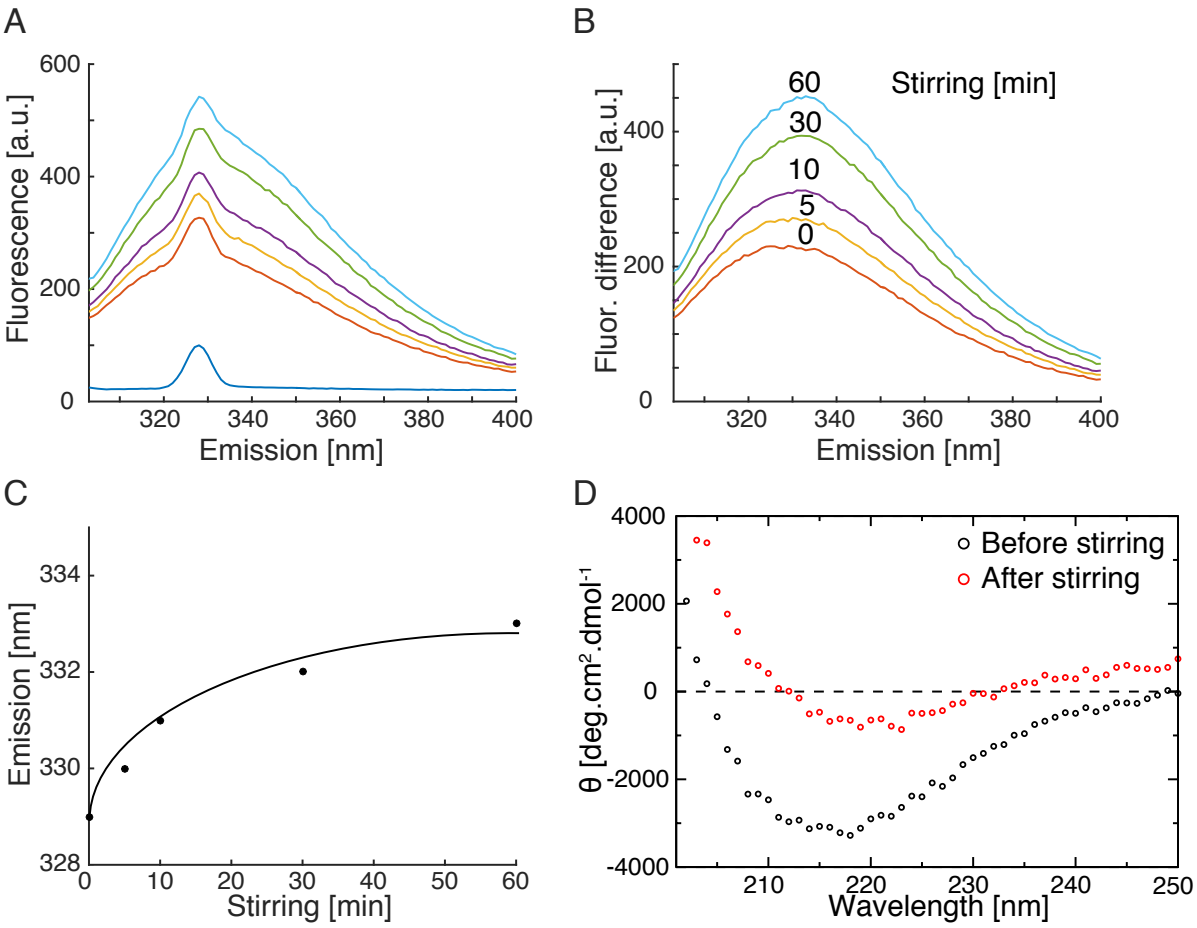

Figure 19: Effect of mechanical stirring on the fluorescence signal and the CD spectrum. **A** Emission spectrum of a 0.5 µM LBP solution stirred over a period of 60 min in a 5 x 5 mm quartz cuvette with a magnetic bar stirrer at about 200 RPM. The line coloring correspond to different time points indicated in subpanel B, with the exception of the blue line at the bottom that is the fluorescence signal of the buffer alone. All details of the fluorescence measurements were as described in Figure 14. **B** Emission spectrum after buffer subtraction. **C** Plot of the integrated fluorescence signal (between 308-318 nm and 335-380 nm) as function of the total stirring time. **D** CD spectrum of LBP recorded (0.2 cm quartz cuvette), before and after stirring and fluorescence measurement.

LBP-LPS interaction probed by surface plasmon resonance (SPR)

LBP immobilization on the SPR sensor chip surface

The use of SPR for studying protein-glycoconjugate interactions has been increasing over the last two decades (62). Here, we used SPR to obtain kinetic information on the binding of LBP to LPS and other glycolipids. First, LBP was covalently coupled to the dextran surface of a gold sensor chip from the SIA KIT AU (GE Healthcare)using primary amines groups (Figure 20). The total amount immobilized was 238.5 RU, which corresponds to an average distance of 199 Å between LBP molecules on the chip surface (see legend to Figure 20).

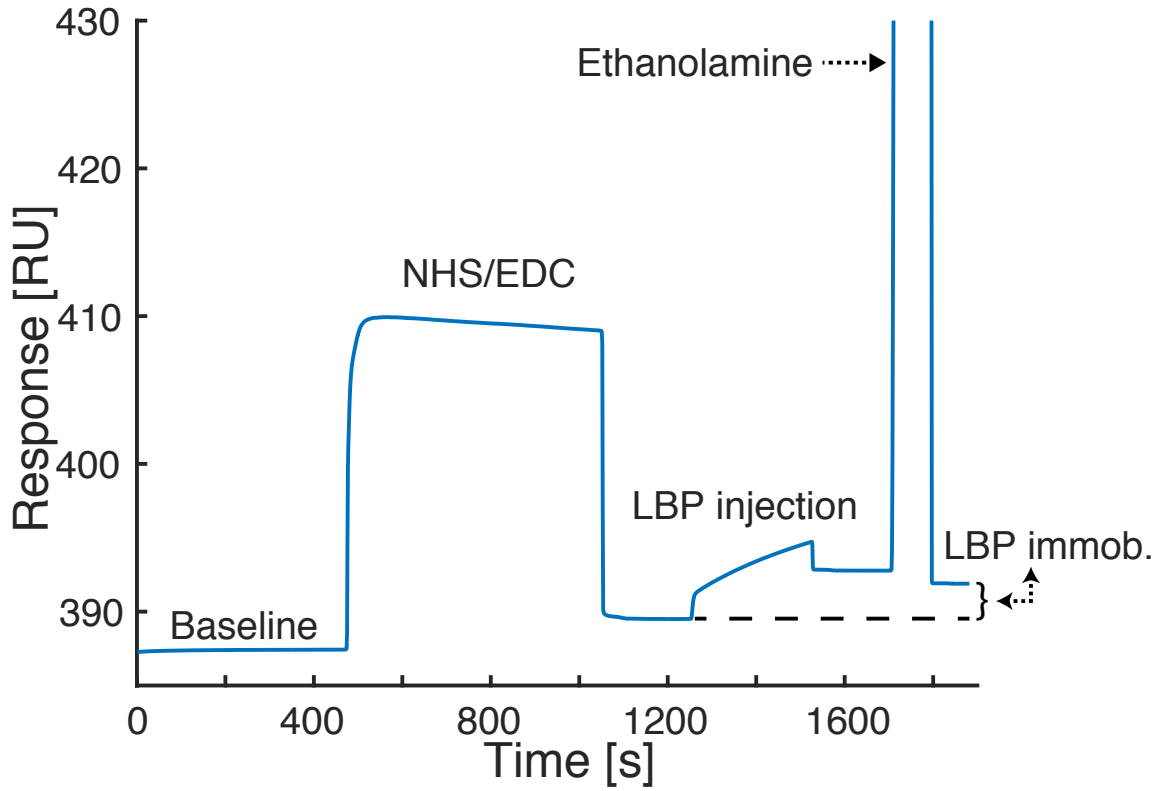

Figure 20: Coupling of LBP to the sensor chip using amine chemistry. The chip surface was activated with NHS/EDC, flushed with LBP and blocked with ethanolamine, at pH 7.5. For coupling, a stock solution of 3.5 µM LBP was injected at 2 µL/min for 300 sec. The LBP immobilized produced a total change in RU of 238.5. The experiment was carried out in 5 mM HEPES, 200 mM NaCl, 1 mM EDTA, pH 7.5. Since 1 RU corresponds to 1 pg/mm², relation between SPR response R and average distance is given by
$R = 238.5\ \text{RU} = \frac{0.2385.10^{-9}\text{g}}{\text{mm}^2}.\frac{\text{mol}}{57.10^3\text{g}} = \frac{2.52.10^{-5}}{\text{Å}^2}$; $\overline{d} = \frac{1}{\sqrt{R}}$=199 Å.

Analysis of the interaction of LBP with LPS and its precursors

The interaction of LBP with the glycolipid ligands resulted in the binding sensorgrams shown in Figure 21. Binding curves were collected for different ligands at different concentrations, and all produced an increase of the SPR response upon titration of increasing concentrations of the different glycolipids, clearly confirming the binding of these molecules to the immobilized protein. The comparison of the sensorgrams for the different ligands showed that the dissociation of Lipid X, although also slow, is faster than the dissociation of Lipid IVa, LPS F515 and LPS EH100. This result indicates that Lipid X has a larger dissociation rate (k_{off}) compared to the other glycolipids. The SPR equilibrium binding experiments (Table 3, discussed below) show that the binding affinity increases in the following order: Lipid X < Lipid IVa < LPS F515 < LPS EH100. Thus the lowest affinity corresponds to the fastest dissociation rate. In contrast, the association rates (k_{on}) could not be determined reliably since simple reaction models did not provide a satisfactory fit.

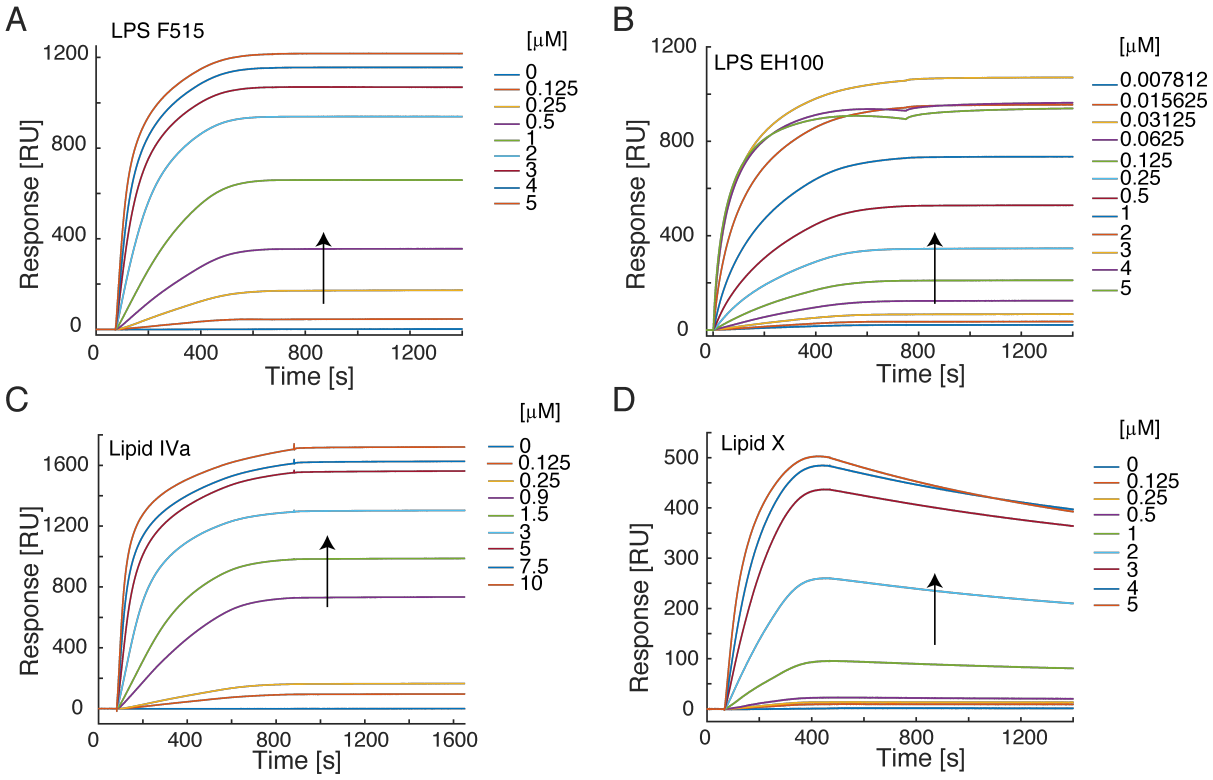

Figure 21: Sensorgrams of the titrations using different glycolipids. **A** 0, 0.125, 0.25, 0.5, 1, 2, 3, 4 and 5 µM of LPS F515, injected over the surface with a contact-time of 400 s **B** 0.01, 0.02, 0.03, 0.06, 0.125, 0.25, 0.5, 1, 2, 3, 4 and 5 µM of LPS EH100, injected over the surface with a contact-time of 750 s. **C** 0, 0.125, 0.25, 0.9, 1.5, 3, 5, 7.5 and 10 µM of Lipid IVa, injected over the surface with a contact-time of 800 s. **D** 0, 0.125, 0.25, 0.5, 1, 2, 3, 4 and 5 µM of Lipid X, injected over the surface with a contact-time of 400 s. All experiments were run at 25 ºC in 5 mM HEPES buffer, 200 mM NaCl, 1 mM EDTA, pH 7.5.

The equilibrium binding could be fitted well by the Hill equation (Eq. 2):

$$R = \frac{R_{max}}{1+(K_d/[L])^n} \qquad \text{(Eq. 2)}$$

where $[L]$ corresponds to the concentration of ligand and n is the Hill coefficient. The fits are shown in Figure 22 and the corresponding fit parameters are given in Table 3.

The equilibrium SPR response (R_{max}) of the different glycolipids follows the trend Lipid IVa > LPS F515 > LPS EH100 > Lipid X > GlcN(3-OH-14:0)-α-P. In the concentration range from 0 to 10 µM, only glycolipids with more than one acyl chain (Lipid IVa, LPS F515, LPS EH100 and Lipid X) produced an SPR response (Figure 22), but the single lipid chain GlcN(3-OH-14:0)-α-P did not show any binding. As shown below, however, GlcN(3-OH-14:0)-α-P produced an SPR response at 170 µM (Figure 24), which is close to its critical micelle concentration (CMC).

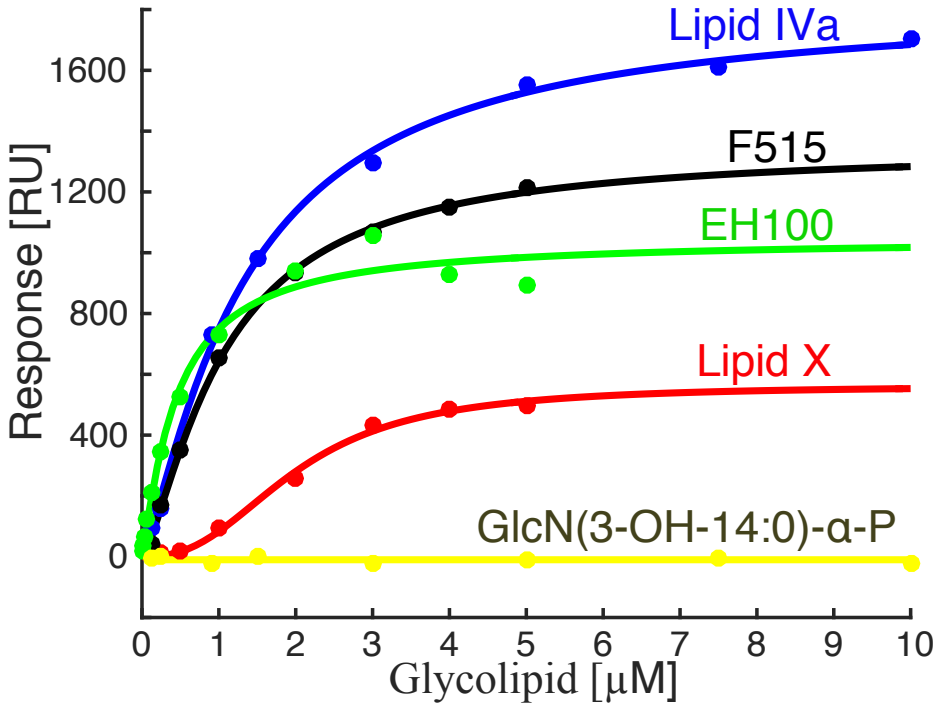

Figure 22: Plot of the equilibrium SPR responses (filled circles) for the different glycolipids versus their concentrations. The solid lines correspond to a fit of the Hill equation (Eq. 2, see text).

Table 3: Binding data derived from the SPR equilibrium experiments, and aggregate sizes determined by DLS.

	Glycolipid [Da][a]	R_{LBP} [RU][b]	R_{Glyc} [RU][c]	Aggregate [kDa][d]	Glycolipid monomers[e]	K_d [μM][f]	n[g]	CMC_{Lit} [μM][h]	CMC_{DLS} [μM][i]
GlcN(3-OH-14:0)-α-P	485.5	353.9	0.0	-	-	-	-	-	62.5-125.0
Lipid X	711.9	238.5	562.7	134.5	189	2.0	2.5	40[j]	3.9-31.3
Lipid IVa	1405.7	353.9	1821.8	293.4	209	1.3	1.2	< 0.1[j]	-
LPS F515	2237.3	238.5	1343.4	321.1	144	1.1	1.4	~0.04[k]	-
LPS EH100	5500.0	238.5	1045.9	250.0	45	0.5	1.2	> 0.04[k]	-

[a] Molecular weight of the glycolipid monomer.

[b] SPR response produced upon coupling of LBP to the sensor chip surface.

[c] Maximum response produced upon binding of glycolipids to surface-immobilized LBP. Calculated from fit to the Hill equation.

[d] MW of bound glycolipid aggregate per LBP monomer determined as: $MW_{Aggregate} = \frac{MW_{LBP} \times R_{glycolipid}}{R_{LBP}}$ and MW_{LBP} is 57 kDa.

[e] Number of glycolipid monomers forming the interacting aggregate. Obtained from the molecular weights of the aggregate and the monomers.

[f] Equilibrium binding constant derived from Hill equation.

[g] Hill coefficient of cooperativity.

[h] CMC of the glycolipids reported in literature.

[i] Reference (63).

[j] Reference (64).

[k] Reference (65).

[l] CMC of the glycolipids determined experimentally by DLS.

R_{max} is dependent on the surface capacity of the immobilized LBP and the molecular mass of the glycolipids. While saturation means that all the interaction sites have been occupied, the maximum amplitude response gives an indication of the total amount of the glycolipid bound. It is noteworthy that the much smaller R_{max} observed for Lipid X corresponds approximately to the same number of bound glycolipid monomers as Lipid IVa and LPS F515.

The Hill coefficient *n* resulting from the binding of the different glycolipids to the surface-immobilized LBP is a measure of the degree of cooperativeness of these interactions. The values obtained for most are approximately 1, which indicates lack of cooperativity. In contrast, the interaction of LBP with Lipid X was well fitted with a value *n* of 2.5 indicating cooperativity. The midpoint of the binding isotherm occurs at about 2 μM Lipid X, which is close to the determined range of its CMC (3.9 – 31.3 μM, see below). Very likely the observed cooperativity describes the formation of Lipid X micelles, which bind to LBP.

As remarked before, the binding affinity decreases slightly as follows: LPS EH100 > LPS F515 > Lipid IVa > Lipid X. It is remarkable that LPS EH100, closest to full-length LPS, displays the highest affinity towards LBP.

Critical micelle concentration (CMC) of Lipid X and GlcN(3-OH-14:0)-α-P glycolipids

In order to better understand the binding behavior of the Lipid X and GlcN(3-OH-14:0)-α-P, their apparent CMC was determined using the change of aggregation size determined by DLS. For Lipid X, we observe that the transition between small to very large particle size occurs in the concentration range above 3.9 μM and below 31.3 μM (Figure 23A), while for GlcN(3-OH-14:0)-α-P the transition occurs above 62.5 μM and below 125 μM (Figure 23B). This indicates that the corresponding CMC of both glycolipids lies in those ranges. While no CMC values were found in the literature for GlcN(3-OH-14:0)-α-P, the CMC of Lipid X was reported to be 40 μM (63), which is close to the upper limit of 31.25 μM found in this study.

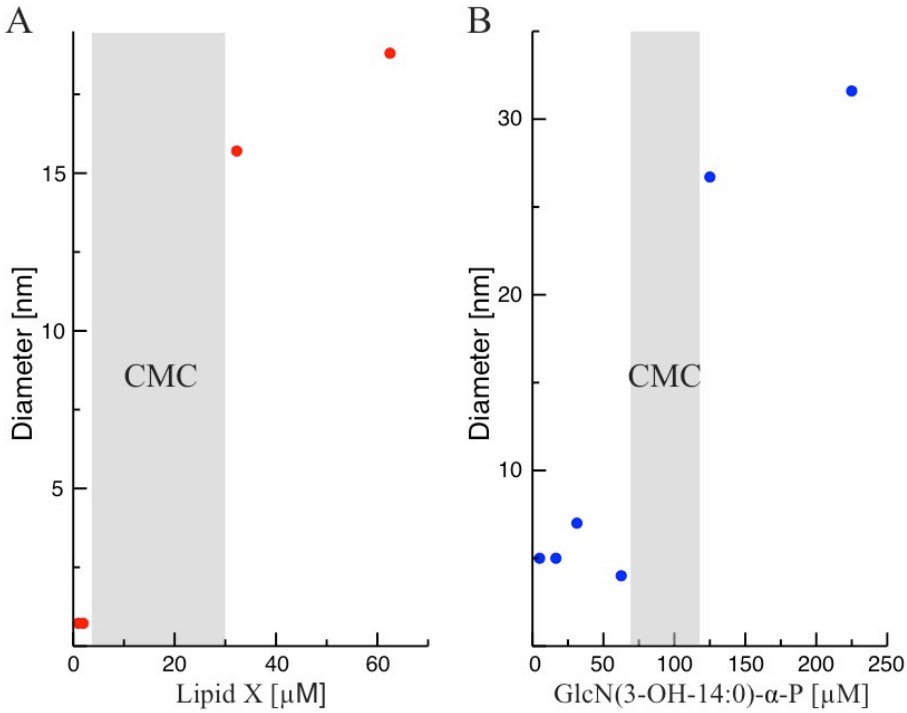

Figure 23: Determination of the CMC of Lipid X and GlcN(3-OH-14:0)-α-P. **A** Measured size of Lipid X at the concentrations of 0.98, 1.95, 3.90, 31.25 and 62.5 µM. **B** Size of GlcN(3-OH-14:0)-α-P at the concentrations in of 5, 16.5, 31.25, 62.5, 125 and 225 µM. The grey box indicates the concentration range were the transition from monomers to micelles takes place. All experiments were carried out in 5 mM HEPES, 200 mM NaCl, 1 mM EDTA, pH 7.5.

As described before, the GlcN(3-OH-14:0)-α-P does not produce an SPR response in the concentration range of 0-10 µM (Figure 22). However, when the concentration was increased to 170 µM, an SPR response was observed (Figure 24). This indicates that at higher concentrations, GlcN(3-OH-14:0)-α-P micelles bound to LBP produce a large enough SPR response. At concentrations below the CMC, however, the detection limit imposed by the Biacore instrument makes it difficult to detect a clear binding of monomers.

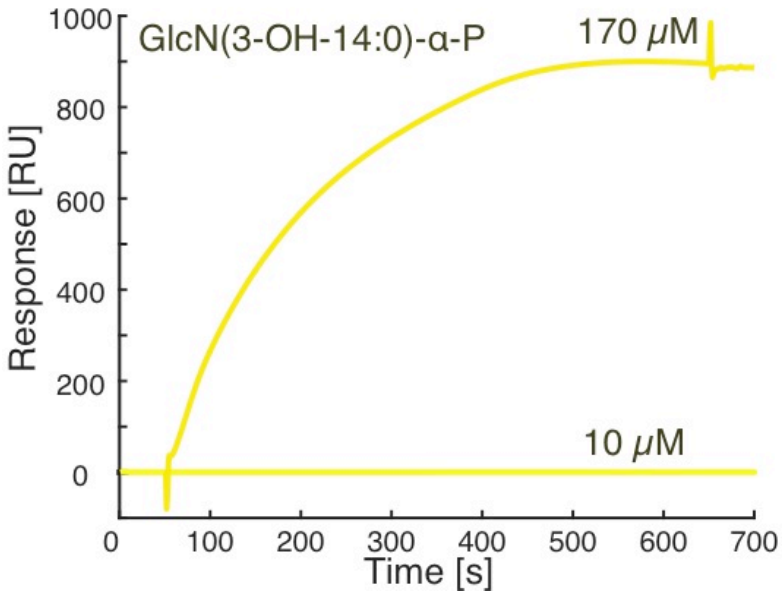

Figure 24: SPR response produced upon injection of 10 µM and 170 µM GlcN(3-OH-α-P) over the surface with a contact-time of 650 s.

Investigation of the importance of electrostatic interactions on the binding

The effect of electrostatics on the LBP-glycolipid binding observed in the fluorescence experiments was also verified by SPR (Figure 25). The SPR response upon binding of Lipid X to the surface-immobilized LBP is continuously reduced in buffers of increasing concentrations of NaCl. In fact, as judged by the maximal SPR response, the total bound lipid mass per LBP molecule is greatly reduced from 165 to 67 kDa (Table 4). On the other hand, a comparison of the inflection points of the binding isotherms indicates that the K_d at 0.2 and 1.0 M NaCl is approximately 2 and 3 µM, respectively; this suggest only a slightly reduced affinity at higher ionic strengths. This could result from either smaller micelles being formed or a reduced number of surface-immobilized LBP molecules interacting with these micelles.

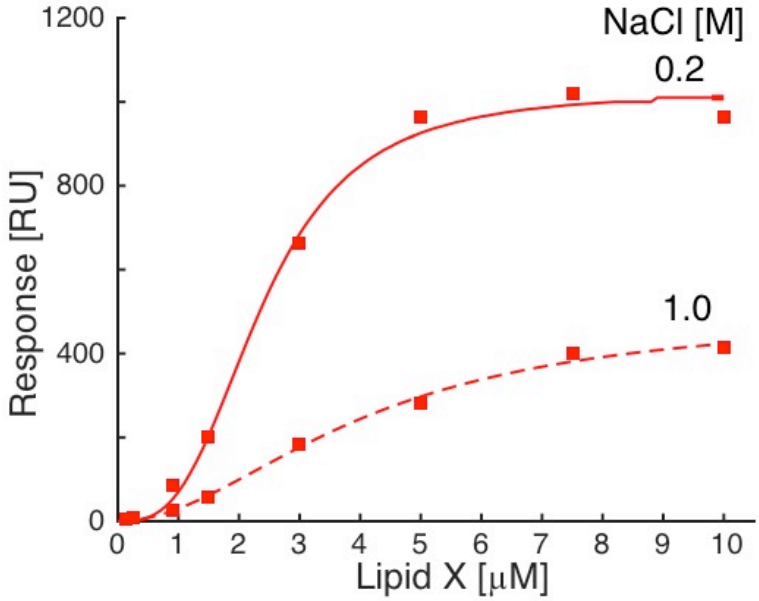

Figure 25: Effect of varying the NaCl concentration on the equilibrium SPR response.

Table 4: Effect of electrostatics on the binding of LBP to Lipid X micelles.

NaCl [M]	R_{LBP} [RU][a]	$R_{Lipid\ X}$ [RU][b]	MW_{agg} [kDa][c]	Glycolipid monomers[d]
0.2	353.9	1025	165	232
1.0	353.9	415	67	94

[a] SPR response produced upon coupling LBP to the sensor chip surface.
[b] Maximum SPR response produced upon binding of Lipid X to the sensor-immobilized LBP.
[c] MW of the bound glycolipid aggregate per surface-immobilized LBP molecule.
[d] Number of bound glycolipid monomers per immobilized LBP molecule.

Characterization of the LBP-LPS interaction

The following experiments were carried out with the purpose to better understand (i) the stoichiometry of the LBP-LPS interaction and (ii) the role of CD14 for the LBP-LPS interaction.

After immobilizing LBP on the SPR surface, 100 nM of LPS F515 was injected, producing an SPR response of 67.5 RU. This response level is about twenty-fold smaller than the one obtained upon saturation of the surface-immobilized LBP (Table 3). Since the saturation experiment (Figure 22) showed no cooperativity, it can be assumed that the micelle sizes are similar. Therefore only a fraction of the LBP molecules are "occupied" with these micelles. The total response of 67.5 RU upon binding of LPS F515 indicates that 7.3 glycolipid monomers are bound per total immobilized LBP.

Subsequently, either LBP (Figure 26B) or CD14 (Figure 26C) was injected over the immobilized LBP-LPS complex. Flowing free LBP produced a response corresponding to the binding of 1.1 LBP molecules per LBP immobilized. This corresponds to 7.3 LPS molecules per 2.1 (1.1 + 1) LBP molecules, in other words a ratio of ~ 3:1. This important result indicates that many LBP molecules can cover/coat the surface of LPS aggregates almost completely.

The addition of CD14 (50 kDa) did not produce a significant response (Figure 26C) on the immobilized LBP-LPS complex. This result indicates that CD14 either does not bind or possesses a very slow association rate.

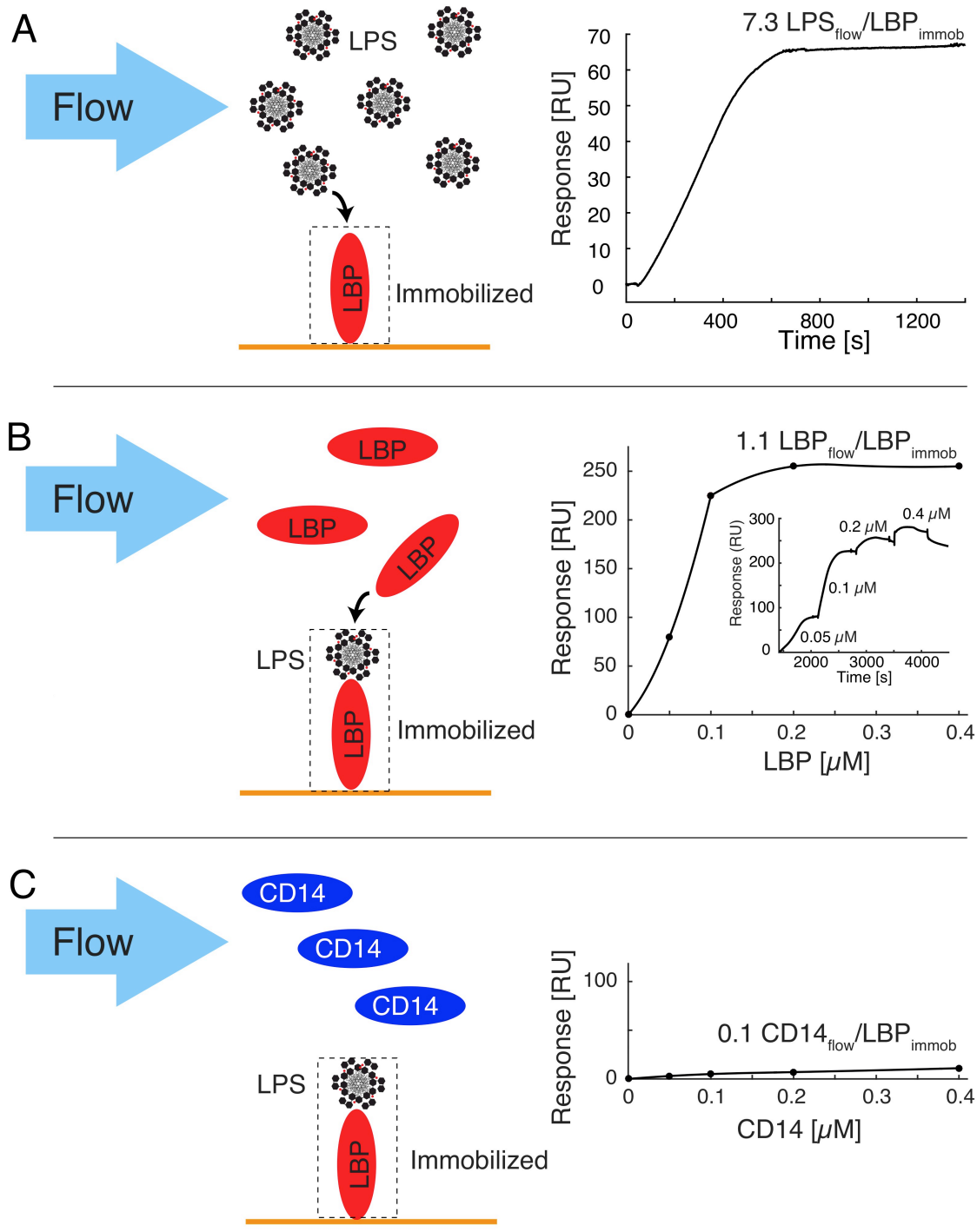

Figure 26: Characterization of the LBP-LPS-CD14 interaction. **A** Scheme for the experimental setup for the immobilization of 100 nM LPS F515 (67.5 RU) on the chip-immobilized LBP (238.5 RU). The total SPR response upon interaction corresponds to the binding of 7.1 LPS monomers per molecule of surface-immobilized LBP. **B** Scheme of the injection of free LBP on the immobilized LPS produces a response of 255 RU. This response corresponds to the binding of 1.1 molecules of free LBP per molecule of surface-immobilized LBP. **C** Injection of free CD14 on the immobilized LPS F515 produces a response of 12 RU, consistent with the binding of 0.1 molecules of free CD14 per molecule of surface-immobilized LBP.

LBP-induced disaggregation of glycolipids requires CD14

After immobilization of LBP on the surface of the SPR chip, the effect of injecting either glycolipids alone (Lipid X, Lipid IVa, LPS F515; Figure 27A) or in combination with CD14 (Figure 27B) was investigated. The results indicate that injecting a mixture of glycolipids and CD14 produces on average only half of the SPR response as compared to injecting the glycolipids alone (Figure 27C). This seems to indicate the onset of disaggregation. The data suggest that disaggregation may require an initial transient interaction between LBP and CD14, which can only be accomplished when LPS and CD14 are injected simultaneously. This hypothesis would explain the absence of micelle disaggregation when CD14 is injected after LPS has been immobilized as previously shown in Figure 26C.

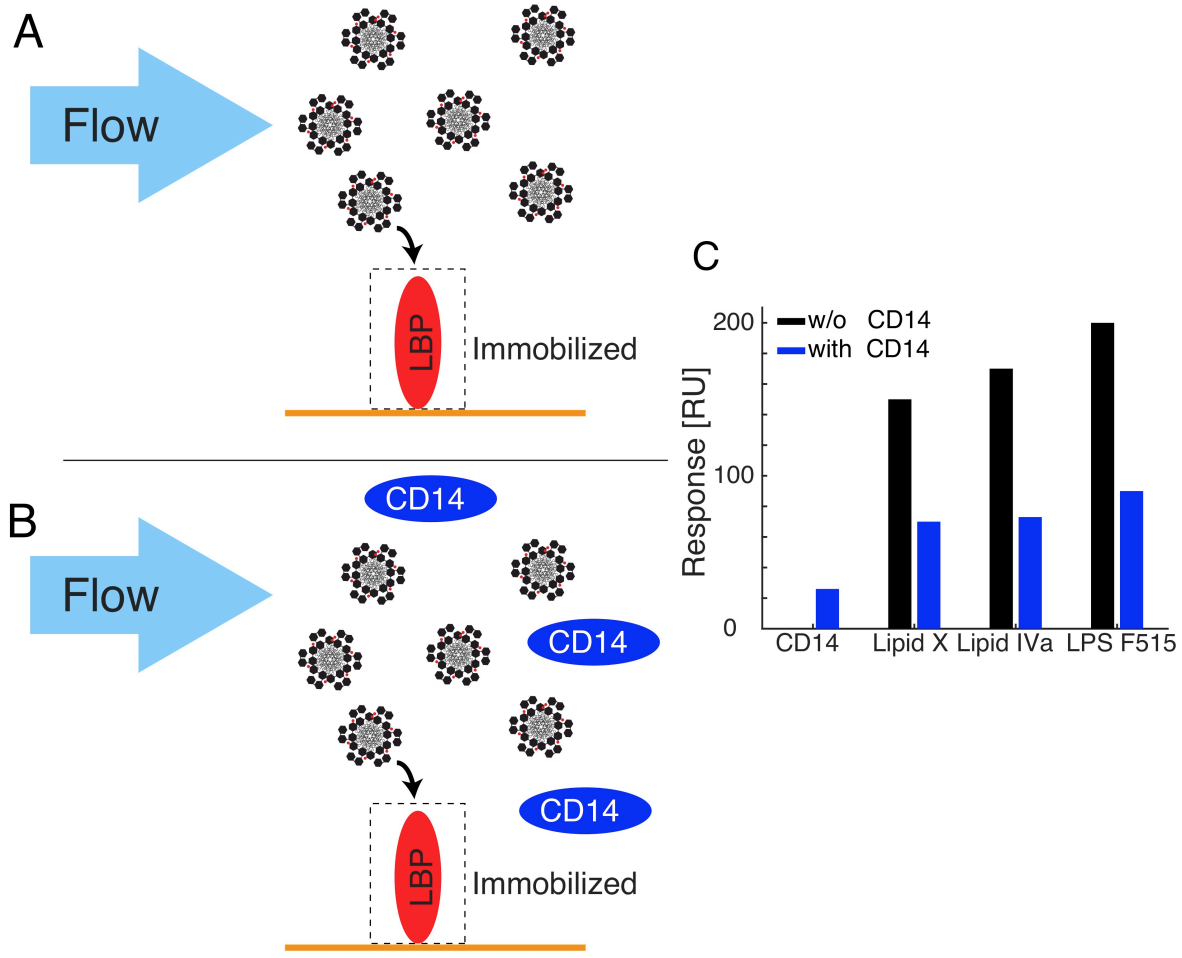

Figure 27: Comparison of the interaction of LBP and glycolipids, in presence or absence of CD14. **A** Scheme of injecting glycolipid over surface-immobilized LBP. **B** Scheme of injecting a mixture of glycolipid and CD14 over surface-immobilized LBP. **C** Bar plot showing that the response produced upon injection of the mixture (blue) of glycolipid (100 nM) and CD14 (100 nM) is half of that produced when injecting the glycolipids (100 nM) alone (black).

Conclusions

In this study, we performed large-scale expression and purification of the *h*LBP from insect cells. We characterized this material using several biophysical methods and showed that it is an ~57 kDa folded glycoprotein containing a mixture of α-helices and β-sheets similar to that found for the *m*LBP. Furthermore, we developed a labeling scheme using a combination of AE and the homemade YE that allowed ~80 % ^{15}N isotope incorporation. A ^{15}N-TROSY HSQC spectrum was for the first time reported for any of the receptors that bind to LPS. The spectrum showed 324 backbone resonances out of 473 expected. However, the spectral quality can be further improved by deuteration as has been shown by Opitz et al. for other proteins (54). The ^{15}N T_1 and T_2 measurements on the labeled LBP confirmed its monomeric state.

The fluorescence of the two LPB tryptophans Trp^{116} and Trp^{176} located in the N-terminal tip was used as a local probe of their environment. Their emission spectrum is blue-shifted and quenched by a factor of four when compared to free tryptophan in aqueous solution. This indicates that the environment of LBP's two tryptophan residues is more hydrophobic than water and that adjacent residues quench the tryptophan fluorescence in apo LBP. Binding of different glycolipids reduces this quenching indicating local conformational changes in the vicinity of the tryptophans and lipid binding pocket of LBP. Higher fluorescence intensities were observed for the larger glycolipids (LPS F515 and Lipid IVa). Using tryptophan fluorescence we also observed that the binding of LBP to the glycolipids is greatly reduced as the ionic strength of the medium is increased, therefore suggesting an important role of electrostatic interactions.

SPR was used to get a better understanding of the kinetics, affinity and binding stoichiometry of the LBP-glycolipid interactions. Equilibrium binding experiments showed a binding affinity in the low micromolar range for most investigated glycolipids, which increases slightly with their size. For GlcN(3-OH-14:0)-α-P binding was observed only above its CMC at 170 μM and no equilibrium binding could be carried out. Its affinity is apparently the lowest of all investigated glycolipids. The biacyl chain Lipid X is the second weakest binder. Fitting of the data using the Hill model to the equilibrium binding isotherms showed that most of the interactions are non-cooperative ($n \sim 1$). The binding of Lipid X, however, produced an n of 2.5, indicating cooperativity among the glycolipid monomers consistent with aggregate formation. The transition is close to the determined concentration range of its CMC. Thus for both GlcN(3-OH-14:0)-α-P and Lipid X binding was only observed in the aggregated form. Monomeric binding would have been below the detection limit of the SPR experiment.

By SPR, we also showed that LBP can bind to LPS with a stoichiometry of approximately 1:3. This is an interesting result as it indicates that LBP binds multiple LPS's, and suggest that in presence of large aggregates it can coat their surface. However, the disaggregation of the LPS micelle only takes place in the presence of CD14.

Materials and Methods

Generation of expression construct

A synthesized gene sequence of the *h*LBP was purchased from a commercial supplier (Genscript). Recombinant Baculovirus was prepared by insertion of the *h*LBP gene into the modified pACEBAC 1 vector (Geneva Biotech), fused downstream of the HBM secretory sequence (MKFLVNVALVFMVVYISYIYAD), and upstream of the GS linker (GSGSGS) and the ten-residue histidine (HHHHHHHHHH). The final sequence is

MKFLVNVALVFMVVYISYIYADANPGLVARITDKGLQYAAQEGLLALQSELL-
RITLPDFTGDLRIPHVGRGRYEFHSLNIHSCELLHSALRPVPGQGLSLSISDS-
SIRVQGRWKVRKSFFKLQGSFDVSVKGISISVNLLLGSESSGRPTVTASSCSSDIAD-
VEVDMSGDLGWLLNLFHNQIESKFQKVLESRI-
CEMIQKSVSSDLQPYLQTLPVTTEIDSFADIDYSLVEAPRATAQ-
MLEVMFKGEIFHRNHRSPVTLLAAVMSLPEEHNKMVYFAISDYVFNTASLVYHEE-
GYLNFSITDDMIPPDSNIRLTTKSFRPFVPRLARLYPNMNLELQGSVPSAPLLNF-
SPGNLSVDPYMEIDAFVLLPSSSKEPVFRLSVATNVSATLT-
FNTSKITGFLKPGKVKVELKESKVGLFNAELLEALLNYYILNTFYPKFND-
KLAEGFPLPLLKRVQLYDLGLQIHKDFLFLGANVQYMRVGSGSGSHHHHHHHHHH

where the blue residues correspond to the mature *h*LBP sequence (UniprotKB entry P18428).

Preparation of recombinant baculovirus

Handling and maintenance of Sf9 cell culture was based on established protocols (66). Permanent cultures were subcultured in mid-log phase and kept in full SF4 medium (Biochem Biocon-cept) Insect Xpress medium (Lonza) at 27 °C and shaken at 80 rpm. Recombinant bacmid DNA was generated, according to supplier's protocol, in *E. coli* DH10Bac cells using a modified pACEBAC 1 (Geneva Biotec) compatible with the Gateway technology (Invitrogen). Baculovirus was generated in adherent cultures as reported by O'Reilly et al. (66). Subsequently, high-titer virus stocks for expression were produced by two additional amplification rounds in suspension cultures. In the first amplification round, 50 mL of Sf9 cell culture containing 1 x 10^6 cells/mL in serum-free medium were infected with recombinant virus and cultured until cell viability decreased to ≤90 %. The supernatant (P1) was collected by centrifugation at 1000g and 4 °C for 15 min. For the second amplification round, 100 mL of serum-free Sf9 cell culture containing 1 x 10^6 cells/mL was infected with P1 and processed as before to yield a P2 virus stock.

Functional titration of the P2 virus stock was applied to determine the optimal amount for the expression. Briefly, small-scale cultures of Sf9 cells were infected in full SF-4 medium at a cell density of 1.5 x 10^6 cells/mL with P2 virus stock corresponding to functional titers of

1-40 mL per 1 L of cell culture. The yield of purified protein was then quantified for different expression times.

Expression and purification of the insect cell-produced hLBP

For large-scale expression, SF-4 or Insect Xpress-adapted Sf9 cells were grown to a density of 2 x 10^6 cells/mL and inoculated with a P2 virus titer of 10 mL per 1 L of cell culture. *h*LBP was subsequently expressed until the cell viability decreased to 80-85 %, usually between 60-72 hours post-infection.

Following expression, medium containing secreted *h*LBP was separated from the cells by centrifugation at 1000*g* for 10 min (4 °C). The supernatant was filtered through a 0.22 µm filter and directly loaded on a strong cation exchange column (SP-Sepharose Fast Flow 2 × 5 mL, GE Healthcare Life Sciences) equilibrated with 25 mM HEPES (pH 7.5), 100 mM NaCl and 5 % glycerol (buffer A). Following a washing step with 2 column volumes (CV) of buffer A, *h*LBP was eluted using a gradient of 25 mM HEPES (pH 7.5), 1 M NaCl and 5 % glycerol (buffer B) from 0-100 % over 20 CV. Fractions containing *h*LBP were combined and subsequently loaded onto a Ni Sepharose column (HisTrap HP 5 ml, GE Healthcare Life Sciences) initially equilibrated with 25 mM HEPES (pH 7.5), 500 mM NaCl and 5% glycerol. The column was washed with 2 CV of buffer A, followed by gradient from 0-5 % in 4 CV of 25 mM HEPES (pH 7.5), 500 mM NaCl, 500 mM imidazole and 5 % glycerol. Bound *h*LBP was finally eluted by a gradient of buffer B from 5-100 % in a total volume of 4 CV. Fractions containing *h*LBP were combined and concentrated by centrifugation in a 30 kDa MWCO filter (VivaSpin 15, Sartorius). The concentrated supernatant was loaded onto a size-exclusion column (Superdex 200 10/300 GL, GE Healthcare Life Sciences) equilibrated with 5 mM HEPES (pH 7.5), 500 mM NaCl and 1 mM EDTA (buffer E). Fractions containing *h*LBP were combined and concentrated for further analysis (VivaSpin 15, Sartorius).

Gel electrophoresis and Western blotting

Protein electrophoresis was performed on precast 4-20 % gradient gels (Bio-Rad) according to manufacturer's protocol. Samples were mixed with 2x non-reducing SDS sample buffer before loading, but not heated. The staining was carried out with the standard coomassie brilliant blue (CBB) procedure.

For Western blots, the protein was transferred to nitrocellulose membranes using a Trans-Blot Turbo system (Bio-Rad). The transfer time was 3 min at 1.3 A and 25 V for a Mini-PROTEAN TGX gel (Bio-Rad). The membrane was blocked for 1 h in 3 % BSA in Tris-buffered saline with 0.1 % Tween-20 (TBST). Subsequently, the membrane was incubated (30 min) with a 1:8000 dilution of primary monoclonal antibody anti-human LBP biG 42 (Biometec), previously prepared as described by manufacturer, in TBST supplemented with 3 % BSA. This was followed by incubation (30 min) with a 1:4000 dilution of horserad-

ish peroxidase (HRP)-conjugated secondary antibody, previously prepared according to manufacturer) in TBST. Washing was performed for 3 x 5 min after incubation with the primary and secondary antibody using TBST. Protein bands were visualized using a chemiluminescent HRP substrate (Roche) according to manufacturer's instructions.

Size exclusion chromatography and static light scattering

Purified *h*LBP (70 µL) was applied to an equilibrated Superdex 200 10/300 GL size exclusion chromatography column. Buffer E was used for equilibration of the column as well as for the experimental run. Fractions were analyzed by UV and static light scattering (Wyatt Technology Corporation).

CD spectroscopy

For analysis of the *h*LBP secondary structure composition, CD spectra of purified *h*LBP were recorded on a Chirascan CD spectrometer (Applied Photophysics Ltd.). The protein (5 µM) was in 5 mM HEPES (pH 7.5), 500 mM NaF and 1 mM EDTA (Buffer F). Measurements were performed in a 0.2 cm quartz cuvette at 25 °C. A sampling of 7 s/nm was used and data recorded in a range of 185-260 nm. Data was corrected for the buffer contribution and analyzed online using DichroWeb (49, 50).

NMR spectroscopy

NMR experiments were recorded on a Bruker Avance III 900 MHz spectrometer equipped with a triple-resonance cryoprobe. All experiments on LBP were acquired at a temperature of 37 °C. Amide proton T_2 values were determined from a jump-return spin-echo experiment (52). The 2D ^{1}H-^{15}N-TROSY experiment on ^{15}N-*h*LBP was recorded with 130 (^{15}N) x 1024 (^{1}H) complex points, acquisition times of 26 ms (^{15}N) and 43 ms (^{1}H), with a total experimental time of 60 h. INEPT transfer times were optimized for amide proton T_2 values of *h*LBP. NMR data were processed and analyzed with the NMRPipe software package (67).

Preparation of glycolipid samples

1 mg of LPS/Lipid A (Na^+, K^+, Ca^{2+}, Mg^{2+}, etc. mixed salt forms) was dissolved in 1.5 mL distilled water and ultra-sonicated for 5 min and kept on ice. Subsequentlly, 2-5 µL of 5 % triethylamine (TEN, Sigma) was added until the opalescence of the sample disappeared (~pH 9). Special care was taken to avoid pH > 9.5, where ester-bound fatty acids are saponified and hydrolyzed. The sample was ultra-sonicated under temperature control and immediately cooled down on ice. After 10-15 min, 150 µL of 1 M HCl (pre-cooled) was added and the sample was shaken to better mix the components. At this point, the Lipid A dispersion became more opalescent as a sign of aggregation. The sample was centrifuged at 13,000 rpm/min for 20 min at 4 °C. The supernatant was discarded. The sediment contains the

LPS/lipid A in its water-insoluble H^+-form. At this point, the desired amount of distilled water (ca. 0.5-1 mL) was added to adjust the final concentration. The sample was kept on ice. Following, 1-5 μL of 20 % TEN was added to achieve pH ~ 7-8. The maximum "solubility" of the Lipid A-TEN-form salt in water is 4-5 mg/mL (~3 mM). This mixture appeared still slightly opalescent but it represents the best water-soluble form of lipid A or LPS. Stock solutions of ~0.5 mg/mL were prepared, aliquoted and frozen at -20 °C for further studies using fluorescence and SPR.

The glycolipids GlcN(3-OH-14:0)-α-P, Lipid IVa (compound 406) (68, 69), Lipid X and Re-like LPS(*E.coli* F515) were kindly provided by Professor Ulrich Zähringer.

Surface plasmon resonance

A commercial plain gold sensor chip from the SIA Kit AU (GE Healthcare) was coated with a mixture containing 200 µL of 1 mM carboxy-thiol (CT)-PEG and 100 µL of 1 mM methyl-thiol (MT)-PEG. The mixture was injected over the SPR chip surface at 2 µL/min for a total of 2400 s, at 25 ºC.

For attaching LBP, the PEGylated SPR chip surface was first activated with a mixture of N-hydroxysuccinimide and N-ethyl-N'-(dimethylaminopropyl)carbodiimide. Subsequently, LBP was captured by flowing 3.5 µM of LBP (in 5 mM HEPES, pH 7.4, 200 mM NaCl, 1 mM EDTA) over the surface at 2 µL/min for 300 s. The unreacted NHS moieties on the surface were blocked with ethanolamine.

For analysis of the biospecific interaction, glycolipids were injected over the surface at varying concentrations and at a flow-rate of 10 µl/min, 25 ºC. The surfaces were regenerated by a 60 s flow of 40 mM β-octyl glucoside (β-OG) at 10 µl/min, and then followed by an injection of 5 M NaCl at 10 µL/min. All measurements were carried out in 5 mM HEPES, pH 7.4, 200 mM NaCl, 1 mM EDTA.

Table 5 shows the experimental difficulties in reproducing the absolute size of the SPR signal in repeated experiments with the glycolipid-LBP system.

Table 5: Variation of the absolute size of the SPR signal in repeated experiments with the glycolipid-LBP system. Responses generated by injection of either 100 or 125 nM F515 LPS on the immobilized LBP, and the resulting binding stoichiometry.

Experiment	R_{LBP} [RU][a]	LPS F515 [nM]	R_{LPS} [RU][b]	LPS F515 Monomers[c]
1	353.9	125	25	2
2	238.5	125	50	5
3	238.5	100	200	21
4	238.5	100	120	13
5	238.5	100	67.5	7
6	238.5	100	75	8
7	238.5	100	70	7

[a] Response upon coupling of LBP to the sensor chip surface.
[b] Response upon injection of LPS F515 on the immobilized LBP.
[c] Number of bound LPS F515 monomers per immobilized LBP.

Tryptophan fluorescence measurements

Fluorescence measurements were carried out on a Jasco fluorescence spectrophotometer using a 5x5 mm quartz fluorescence cuvette. The excitation wavelength chosen for all spectra was 295 nm with a slit width of 5 mm and an integration time of 0.05 s. The background intensity of the samples without protein was recorded. Titration experiments were performed by pipetting different aliquots of glycolipids suspended in HEPES buffer into the cuvette containing 0.25 or 0.5 μM LBP dissolved into the same buffer. The temperature of the cuvette was maintained at 25 °C.

Mass spectrometry

Identification of *h*LBP glycosylation sites was carried out by the in-house MS facility using PNGase F (New England Biolabs) cleavage.

References

1. Janeway CA, Travers P, Walport M, Shlomchik M (2001) *Immunobiology: the immune system in health and disease. 5th edition.* (Garland Science, New York).

2. Chu H, Mazmanian SK (2013) Innate immune recognition of the microbiota promotes host-microbial symbiosis. *Nat Immunol* 14(7):668–675.

3. Qiao Q, Wu H (2015) Supramolecular organizing centers (SMOCs) as signaling machines in innate immune activation. *Sci China Life Sci* 58(11):1067–1072.

4. Iwasaki A, Medzhitov R (2015) Control of adaptive immunity by the innate immune system. *Nat Immunol* 16(4):343–353.

5. Gay NJ, Symmons MF, Gangloff M, Bryant CE (2014) Assembly and localization of Toll-like receptor signalling complexes. *Nat Rev Immunol* 14(8):546–558.

6. Carpenter S, Ricci EP, Mercier BC, Moore MJ, Fitzgerald KA (2014) Post-transcriptional regulation of gene expression in innate immunity. *Nat Rev Immunol* 14(6):361–376.

7. Cao X (2016) Self-regulation and cross-regulation of pattern-recognition receptor signalling in health and disease. *Nat Rev Immunol* 16(1):35–50.

8. Cardoso LS, et al. (2007) Polymyxin B as inhibitor of LPS contamination of Schistosoma mansoni recombinant proteins in human cytokine analysis. *Microb Cell Fact* 6:1.

9. Alexander C, Rietschel ET (2001) Bacterial lipopolysaccharides and innate immunity. *J Endotoxin Res* 7(3):167–202.

10. Zhang H, Niesel DW, Peterson JW, Klimpel GR (1998) Lipoprotein release by bacteria: potential factor in bacterial pathogenesis. *Infect Immun* 66(11):5196–5201.

11. Kirikae T, et al. (1998) Biological characterization of endotoxins released from antibiotic-treated Pseudomonas aeruginosa and Escherichia coli. *Antimicrob Agents Chemother* 42(5):1015–1021.

12. Leeson MC, Morrison DC (1994) Induction of proinflammatory responses in human monocytes by particulate and soluble forms of lipopolysaccharide. *Shock* 2(4):235.

13. Fensom AH, Gray GW (1969) The chemical composition of the lipopolysaccharide of Pseudomonas aeruginosa. *Biochem J* 114(2):185-196.

14. Wang W, Sass HJ, Zähringer U, Grzesiek S (2008) Structure and dynamics of 13C,15N-labeled lipopolysaccharides in a membrane mimetic. *Angew Chem Int Ed Engl* 47(51):9870–9874.

15. Kayagaki N, et al. (2013) Noncanonical inflammasome activation by intracellular LPS independent of TLR4. *Science* 341(6151):1246–1249.

16. Salomao R, et al. (2012) Bacterial sensing, cell signaling, and modulation of the immune response during sepsis. *Shock* 38(3):227–242.

17. Calandra T (2001) Pathogenesis of septic shock: implications for prevention and treatment. *J Chemother* 13 Spec No 1(1):173–180.

18. Brandenburg K, et al. (1993) Influence of the supramolecular structure of free lipid A

on its biological activity. *Eur J Biochem* 218(2):555–563.

19. Seydel U, Oikawa M, Fukase K, Kusumoto S, Brandenburg K (2000) Intrinsic conformation of lipid A is responsible for agonistic and antagonistic activity. *Eur J Biochem* 267(10):3032–3039.

20. Beutler B, Rietschel ET (2003) Innate immune sensing and its roots: the story of endotoxin. *Nat Rev Immunol* 3(2):169–176.

21. Zaehringer U, Lindner B, Rietschel ET (1994) Molecular structure of lipid A, the endotoxic center of bacterial lipopolysaccharides. *Adv Carbohydr Chem Biochem* 50:211–276.

22. Raetz CRH, Reynolds CM, Trent MS, Bishop RE (2007) Lipid A modification systems in gram-negative bacteria. *Annu Rev Biochem* 76:295–329.

23. Molinaro A, et al. (2015) Chemistry of lipid A: at the heart of innate immunity. *Chemistry* 21(2):500–519.

24. Gioannini TL, et al. (2004) Isolation of an endotoxin-MD-2 complex that produces Toll-like receptor 4-dependent cell activation at picomolar concentrations. *Proc Natl Acad Sci USA* 101(12):4186–4191.

25. Shimazu R, et al. (1999) MD-2, a molecule that confers lipopolysaccharide responsiveness on Toll-like receptor 4. *J Exp Med* 189(11):1777–1782.

26. Nijland R, Hofland T, van Strijp JAG (2014) Recognition of LPS by TLR4: potential for anti-inflammatory therapies. *Mar Drugs* 12(7):4260–4273.

27. Galanos C, Freudenberg MA (1993) Bacterial endotoxins: biological properties and mechanisms of action. *Mediators Inflamm* 2(7):S11–6.

28. Tobias PSP, Soldau KK, Ulevitch RJR (1986) Isolation of a lipopolysaccharide-binding acute phase reactant from rabbit serum. *J Exp Med* 164(3):777–793.

29. Schumann RR, et al. (1990) Structure and function of lipopolysaccharide binding protein. *Science* 249(4975):1429–1431.

30. Tobias PSP, Mathison JCJ, Ulevitch RJR (1988) A family of lipopolysaccharide binding proteins involved in responses to gram-negative sepsis. *J Biol Chem* 263(27):13479–13481.

31. Elsbach P (1998) The bactericidal/permeability-increasing protein (BPI) in antibacterial host defense. *J Leukoc Biol* 64(1):14–18.

32. Kirschning CJ, et al. (1997) Similar organization of the lipopolysaccharide-binding protein (LBP) and phospholipid transfer protein (PLTP) genes suggests a common gene family of lipid-binding proteins. *Genomics* 46(3):416–425.

33. Eckert JK, et al. (2013) The crystal structure of lipopolysaccharide binding protein reveals the location of a frequent mutation that impairs innate immunity. *Immunity* 39(4):647–660.

34. Ooi CE, Weiss J, Elsbach P, Frangione B, Mannion B (1987) A 25-kDa NH2-terminal fragment carries all the antibacterial activities of the human neutrophil 60-kDa bactericidal/permeability-increasing protein. *J Biol Chem* 262(31):14891–14894.

35. Beamer LJ (1997) Crystal structure of human BPI and two bound phospholipids at 2.4

angstrom resolution. *Science* 276(5320):1861–1864.

36. Iovine N, Eastvold J, Elsbach P, Weiss JP, Gioannini TL (2002) The carboxyl-terminal domain of closely related endotoxin-binding proteins determines the target of protein-lipopolysaccharide complexes. *J Biol Chem* 277(10):7970–7978.

37. Lamping N, et al. (1996) Effects of site-directed mutagenesis of basic residues (Arg 94, Lys 95, Lys 99) of lipopolysaccharide (LPS)-binding protein on binding and transfer of LPS and subsequent immune cell activation. *J Immunol* 157(10):4648-4656.

38. Ryu JK, et al. (2016) Reconstruction of LPS Transfer Cascade Reveals Structural Determinants within LBP, CD14, and TLR4-MD2 for Efficient LPS Recognition and Transfer. *Immunity* 46(1):38-50.

39. Hailman E, et al. (1994) Lipopolysaccharide (LPS)-binding protein accelerates the binding of LPS to CD14. *J Exp Med* 179(1):269-277.

40. Tobias PS, Soldau K, Gegner JA, Mintz D, Ulevitch RJ (1995) Lipopolysaccharide binding protein-mediated complexation of lipopolysaccharide with soluble CD14. *J Biol Chem* 270(18):10482–10488.

41. Yu B, Wright SD (1996) Catalytic properties of lipopolysaccharide (LPS) binding protein. Transfer of LPS to soluble CD14. *J Biol Chem* 271(8):4100–4105.

42. van Oers MM, Pijlman GP, Vlak JM (2015) Thirty years of baculovirus-insect cell protein expression: from dark horse to mainstream technology. *J Gen Virol* 96(Pt 1):6-23.

43. Kim J-I, et al. (2005) Crystal structure of CD14 and its implications for lipopolysaccharide signaling. *J Biol Chem* 280(12):11347–11351.

44. Shin HJ, et al. (2007) Kinetics of binding of LPS to recombinant CD14, TLR4, and MD-2 proteins. *Mol Cells* 24(1):119–124.

45. Kim HM, et al. (2007) Crystal structure of the TLR4-MD-2 complex with bound endotoxin antagonist Eritoran. *Cell* 130(5):906–917.

46. Park BS, et al. (2009) The structural basis of lipopolysaccharide recognition by the TLR4-MD-2 complex. *Nature* 458(7242):1191–1195.

47. Tessier DC, Thomas DY, Khouri HE, Laliberié F (1991) Enhanced secretion from insect cells of a foreign protein fused to the honeybee melittin signal peptide. *Gene* 98:177-183.

48. Shi X, Jarvis DL (2007) Protein N-glycosylation in the baculovirus-insect cell system. *Curr Drug Targets* 8(10):1116–1125.

49. Whitmore L, Wallace BA (2004) DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. *Nucleic Acids Res* 32:W668–73.

50. Whitmore L, Wallace BA (2008) Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases. *Biopolymers* 89(5):392-400.

51. Touw WG, et al. (2015) A series of PDB-related databanks for everyday needs. *Nucleic Acids Res* 43(Database issue):D364–8.

52. Sklenář V, Bax A (1987) Spin-echo water suppression for the generation of pure-phase two-dimensional NMR spectra. *J Mag Res* 74:469-479.

53. Opitz C, Isogai S, Grzesiek S (2015) An economic approach to efficient isotope labeling in insect cells using homemade 15N-, 13C- and 2H-labeled yeast extracts. *J Biomol NMR* 62(3):373–385.

54. Opitz C (2016) *Doctoral Thesis: Development and application of an economic approach to isotope labeling in higher eukaryotes for NMR studies* (University of Basel).

55. Sitarska A, et al. (2015) Affordable uniform isotope labeling with (2)H, (13)C and (15)N in insect cells. *J Biomol NMR* 62(2):191–197.

56. Fountoulakis M, Lahm HW (1998) Hydrolysis and amino acid composition analysis of proteins. *J Chromatogr A* 826(2):109-34.

57. Farrow NA, et al. (1994) Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by 15N NMR relaxation. *Biochemistry* 33(19):5984-6003.

58. Kay LE, Torchia DA, Bax A (1989) Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry* 28(23):8972–8979.

59. Lakowicz JR (2006) *Principles of fluorescence spectroscopy. Third edition.*

60. Tobias PS, Soldau K, Ulevitch RJ (1989) Identification of a lipid A binding site in the acute phase reactant lipopolysaccharide binding protein. *J Biol Chem* 264(18):10867–10871.

61. Han J, Mathison JC, Ulevitch RJ, Tobias PS (1994) Lipopolysaccharide (LPS) binding protein, truncated at Ile-197, binds LPS but does not transfer LPS to CD14. *J Biol Chem* 269(11):8172–8175.

62. Safina G (2012) Application of surface plasmon resonance for the detection of carbohydrates, glycoconjugates, and measurement of the carbohydrate-specific interactions: a comparison with conventional analytical techniques. A critical review. *Anal Chim Acta* 712:9–29.

63. Lipka G, Hauser H (1990) The phase behaviour of lipid X at higher water concentration. *International Journal of Biochem Cell Biol* 68(1):345–351.

64. Hofer M, Hampton RY, Raetz C, Yu H (1991) Aggregation behavior of lipid IVA in aqueous solutions at physiological pH. 1: Simple buffer solutions. *Chem Phys lipids* 59(2):167-81.

65. Sasaki H, White SH (2008) Aggregation behavior of an ultra-pure lipopolysaccharide that stimulates TLR-4 receptors. *Biophys J* 95(2):986–993.

66. O'Reilly DR, Miller LK, Luckow VA (1994) *Baculovirus expression vectors: a laboratory manual* (The Oxford University Press).

67. Delaglio F, et al. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* 6(3):277–293.

68. Fujimoto Y, et al. (2005) Synthesis of lipid A and its analogues for investigation of the structural basis for their bioactivity. *J Endotoxin Res* 11(6):341-7.

69. Scior T, Alexander C, Zaehringer U (2013) Reviewing and identifying amino acids of human, murine, canine and equine TLR4 / MD-2 receptor complexes conferring endotoxic innate immunity activation by LPS/lipid A, or antagonistic effects by Eritoran, in contrast to species-dependent modulation by lipid IVa. *Comput Struct Biotechnol J* 5:e201302012.

3 Structural and dynamic characterization of the intrinsically disordered hepatitis B virus X protein (HBx)

Abstract

The hepatitis B virus X protein (HBx) is important for viral infection and replication. Despite its multiple assigned cellular roles, the molecular details of its involvement in the pathogenesis of hepatitis B virus (HBV)-induced hepatocellular carcinoma (HCC) remain unclear. The intrinsically disordered nature of HBx and its tendency to aggregate in solution have hindered its structural analysis. In this study, we show that addition of a detergent mixture of foscholine 12 (FOS-12) and 1,2-diheptanoyl-sn-glycerol-3-phosphocholine (DHPC) reduces HBx aggregation, and improves the NMR spectral quality, thus allowing the first complete characterization of the full-length viral protein. The chemical shifts provide evidence of the existence of four regions (α1, α2, α3 and β) with higher secondary structure propensity. The ^{15}N relaxation data of the backbone amide sites indicate that slower local motions characterize these regions, as they show an increase of the ^{15}N R_1, R_2 relaxation rates and $\{^1H\}$-^{15}N NOEs. Remarkably, these four regions overlap with functional motifs described in the literature, suggesting that HBx uses conformational selection with pre-formed secondary structure elements for the recognition of its cellular targets.

Introduction

Despite the existence of vaccines and antiviral treatments, more than 240 million people are still chronically infected with human hepatitis B virus (HBV) and are at a significant risk of developing hepatocellular carcinoma (HCC), the second cause of death from cancer worldwide (1, 2). Thus, there is a strong need to understand the mechanism of HBV replication and persistence.

HBV is a member of the *Hepadnaviridae* family, characterized by a small, enveloped DNA, capable of infecting a limited range of mammals and birds. The HBV genome, with a total length of 3.2 kb, consists of four overlapping open reading frames (ORFs): core, polymerase, surface and X (3).

The smallest ORF X encodes the ~17 kDa HBV protein X (HBx) (4), which is expressed during acute and chronic HBV infections, and is found in both the nucleus and the cytoplasm of the cell (5, 6). Although the functions of HBx are not fully understood, it has been assigned a key role in HBV infection and replication (7-9).

The HBx amino acid sequence comprises several functional sites (Figure 4) for modulating cellular activities. These are: i) transcriptional activation by interactions with transcription factors (10, 11); ii) cell proliferation and viability by interactions with p53 (12), suggesting a link to apoptosis (13, 14); iii) transcriptional transactivation of cellular and viral promoters (15-17).

The most studied HBx interaction partner is the damage-specific DNA binding protein 1 (DDB1) (18-20), a multifunctional protein known for its role in the recognition of UV-damaged DNA (21), and also as a component of the ubiquitin proteasome system (UPS) (22, 23). Evidence for the binding of HBx to DDB1 comes from several observations. The interaction was first demonstrated using the yeast two-hybrid assay (24). HBx contains a conserved binding motif common in other cellular proteins that bind DDB1 (25). Recently, the crystal structure of DDB1 in complex with this motif Ile^{88}-Leu^{100} (Figure 1) (26) has shown that it adopts a three-turn α-helical conformation upon binding to DDB1 at a large pocket enclosed by the β-propeller domains (BPA-BPC). Other studies (27), however, suggest that further structural contacts exist. Thus, the complete structural and dynamic characterization of the full-length HBx is needed for a better understanding of its interactions with different cellular targets.

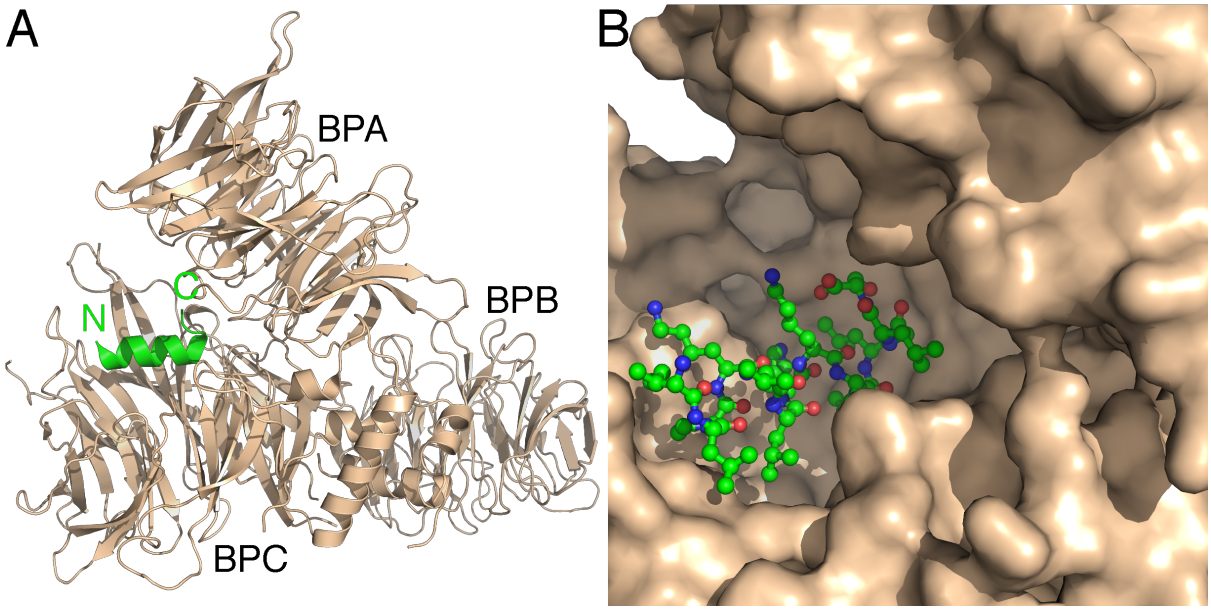

Figure 1: A Structure of DDB1 (wheat, PDB code 3I7H) in complex with the minimal helical fragment (Ile^{88}-Leu^{100}) of HBx (green). **B** Surface representation of DBB1, highlighting the accommodation of the helical fragment in the BPA-BPC double propeller pocket.

So far, an high-resolution structural information on the full-length HBx has been lacking due to the difficulties in producing high yields of the soluble protein (28, 29). The first published 1D NMR spectrum of a truncated mini-HBx(5CS)(18-142) lacked proton signals below 0 ppm (30), and showed a very narrow signal dispersion between 7.0 and 8.5 ppm. Later, a ^{1}H-^{15}N HSQC spectrum of this construct showed only 80 (unassigned) cross peaks instead of the expected 118 (31). These results indicated an unstructured nature for HBx.

In the current study, we set out to determine a more complete structural and dynamic picture of the viral protein. Initial tests showed that a careful selection of a detergent mixture of foscholine 12 (FOS-12) and 1,2-diheptanoyl-sn-glycerol-3-phosphocholine (DHPC) reduces

HBx aggregation. This new condition improves the spectral quality and increases the number of observable backbone ^{1}H-^{15}N resonances. Hence, we were able to assign a total of 113 out of 142 expected backbone cross-peaks and nearly all side-chain atoms. Besides providing the unequivocal confirmation of the unfolded structure, the secondary chemical shifts show that HBx contains 4 regions with transient local structural elements with slower local backbone motions, as indicated by the increase of their ^{15}N R_1, R_2 relaxation rates and $\{^{1}H\}$-^{15}N NOEs. These regions overlap with the important functional sites described in the literature.

Results and Discussion

Backbone and side chain resonance assignments

The reason for the lack of structural information on HBx stems from its inherent tendency to aggregate. In this study, HBx(9CS) was successfully solubilized with a detergent mixture of 2.8 mM FOS-12 and 100 mM DHPC, which increases both the stability and the NMR spectral quality of the viral protein. Hence, the assignment of the backbone and side chain resonances was possible using standard triple-resonance NMR experiments (for details see **Materials and Methods**) on samples of detergent-solubilized ^{15}N,^{13}C-labeled HBx(9CS).

The assignment comprise a total of 113 ^{1}H-^{15}N backbone resonances (Figure 2A), corresponding to 79.6 % of the HBx(9CS) amino acids excluding 11 prolines and the non-native 24 residues (MGSSHHHHHHSSGRENLYFQGSGS) that precede the native protein (full construct shown in **Materials and Methods**). The N-terminal non-native amino acids (decreasing numbering from Ser^{-1} to Met^{-22}) were almost completely assigned, except Met^{-22}, Gly^{-21}, His^{-17}, His^{-13} and Gln^{-3}, which showed severe overlap of their ^{13}C chemical shifts. The ^{1}H-^{15}N HSQC spectrum shows further 28 broad, weak resonances (Figure 2B) that could not be assigned since these were not detected in the 3D experiments. Apparently, this broadening is caused by chemical exchange in the micro- to millisecond time range. These resonances very likely belong to the C-terminal region comprising the amino acid ranges Thr^{105}-Glu^{121} and Leu^{123}-Leu^{134} (29 residues).

Cis-trans isomerism of the proline residues was identified as the cause of conformational heterogeneity at several locations in HBx(9CS). The ^{1}H-^{15}N HSQC spectrum (Figure 2C) shows several examples of the equilibrium between major (trans) and minor (cis) isomers, with population ratios of approximately 80 % and 20 %, respectively. The structural heterogeneities, as indicated by the additional set of resonances, are localized in the vicinity of the prolines. Nevertheless, the global conformations of the different isomers must be very similar due to their similar chemical shifts. In total, we identified 39 residues showing such heterogeneity, with 8 of these having even a third minor population resulting from the vicinity of a second proline residue (Figure S1 in **Supplementary Information**).

Side-chain assignments were obtained from the 3D experiments C(CO)NH and H(CCO)NH (Figure 3A,B), which correlate all the aliphatic ^{1}H or ^{13}C resonances of a given amino acid directly with the amide of the next residue. This approach provided the assignment of 77.5 % of the non-methyl aliphatic side-chains and 60.8 % of the methyl group resonances.

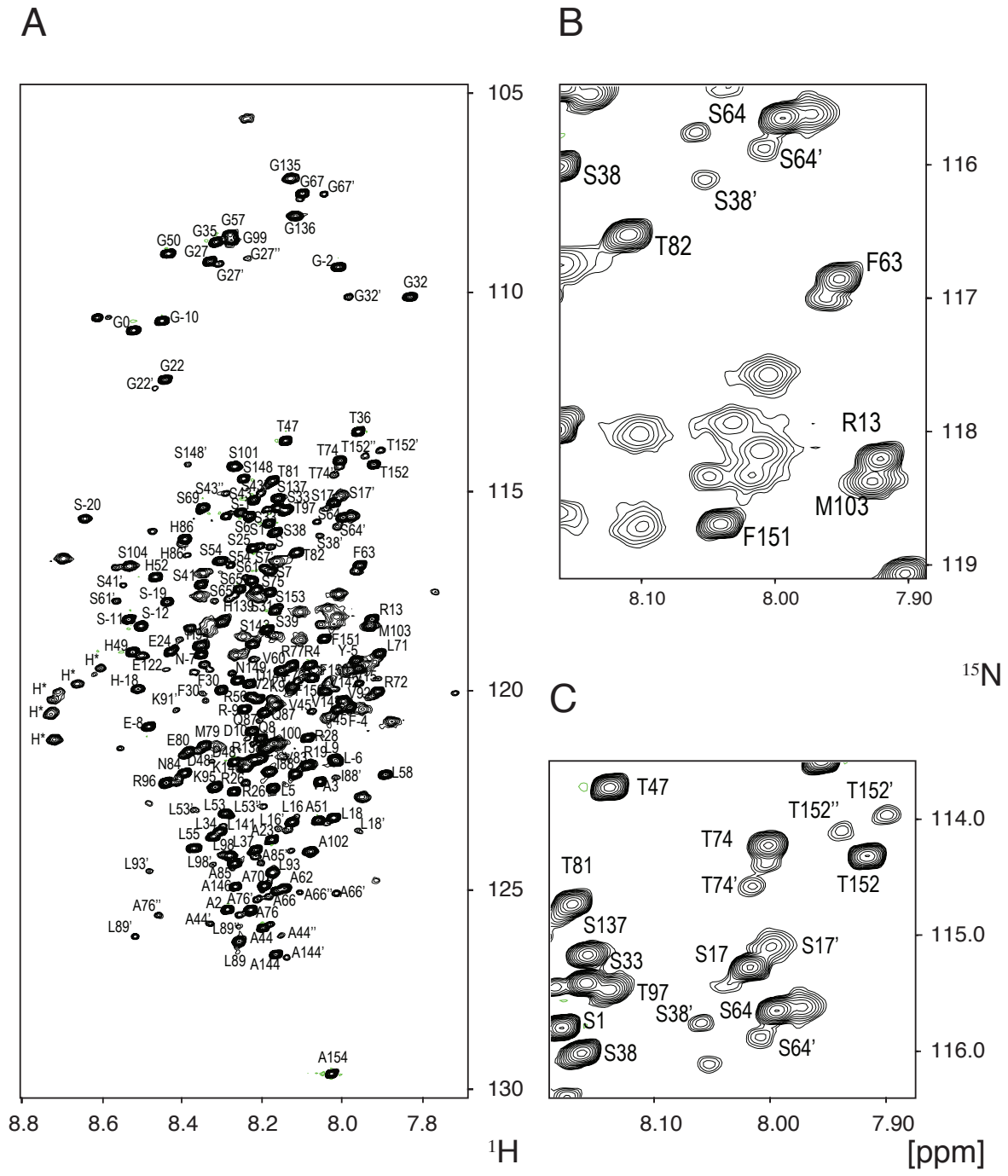

Figure 2: A ^{1}H-^{15}N HSQC spectrum of the uniformly ^{15}N,^{13}C-labeled HBx(9CS), including the assignments of the backbone resonances. Residues Met^{-22}-Ser^{1} belong to the N-terminal non-native sequence. His^{-13}-His^{-17} resonances (marked by asterisks) could not be assigned unequivocally due to overlap of their ^{13}C frequencies. **B** Region of (A) showing some of the residues affected by exchange broadening. **C** Region of (A) showing the splitting of the resonances Ser^{17}, Ser^{38}, Ser^{64}, Thr^{74}, and Thr^{152} due to cis-trans isomerization of the adjacent prolines Pro^{39}, Pro^{40}, Pro^{68} and Pro^{147}.

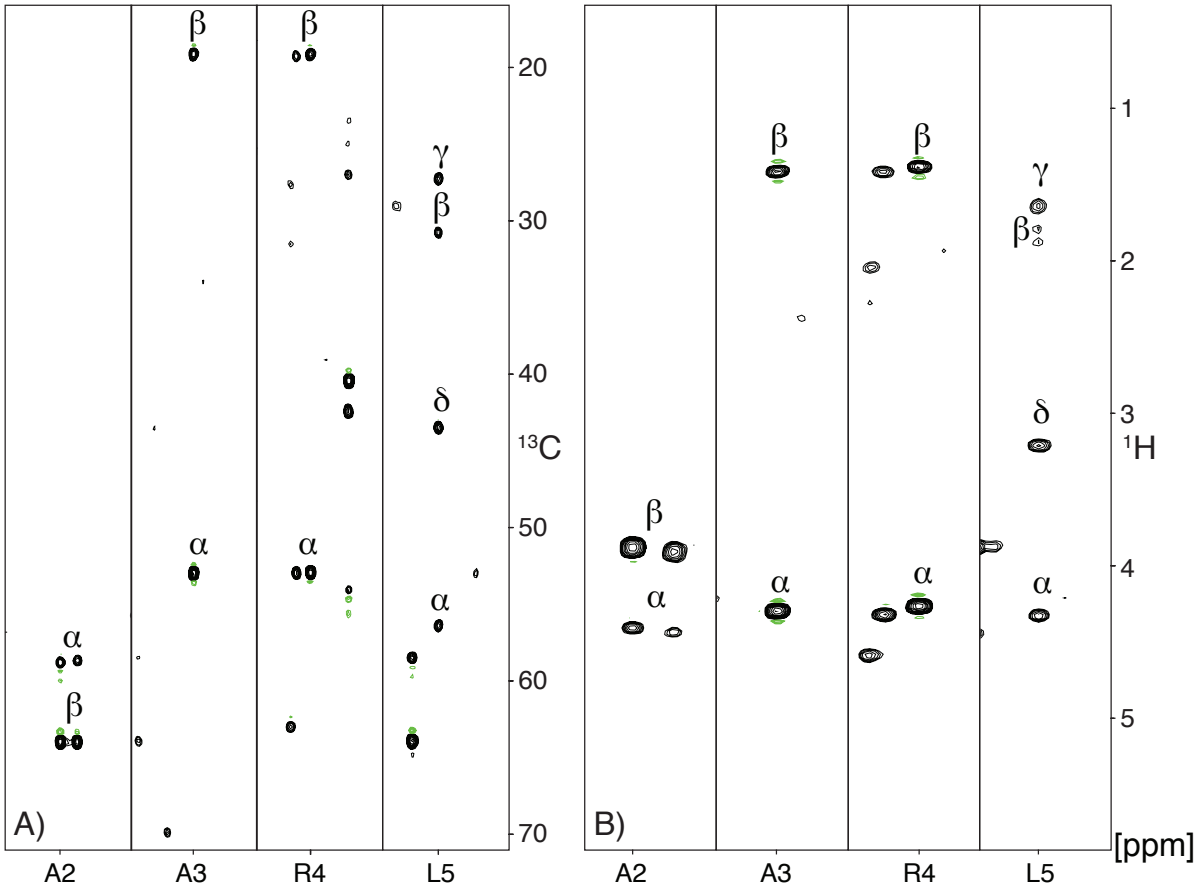

Figure 3: Strip plots of correlations observed for the amides of residues A[2] to L[5] of HBx(9CS) taken from **A** C(CO)NH and **B** H(CCO)NH. Each amide correlates with the aliphatic side-chain carbon and proton frequencies of the preceding residue, respectively.

Secondary structure propensities

The chemical shifts of nuclei are very sensitive probes of the local chemical structure. The secondary chemical shifts, determined as the difference between the experimentally measured and the chemical shifts of random-coil structures, properly corrected for the primary sequence effect, indicate whether a protein segment attains, even if only for a fraction of time, some secondary structure. Relative secondary structure propensities can be estimated from the ratio of measured $^{13}C^{\alpha}$ secondary shifts relative to secondary shift averages of fully formed α-helices (+2.6 ppm) and β-sheets (-1.4 ppm), from the Biological Magnetic Resonance Databank (BMRB).

The positive C^{α} and C', and the negative ^{15}N secondary shifts shown in Figure 4 are consistent with a higher α-helical propensity. Therefore, we defined the regions Ala^{2}-Arg^{28} as α1, Arg^{56}-Thr^{105} as α2 and Leu^{134}-Ala^{146} as α3. As judged by their $^{13}C^{\alpha}$ secondary shifts, these regions correspond to a propensity of about 19 % α-helical structure. Noteworthy, the defined α2 region coincides with the fragment of HBx that binds to DDB1 (26).

The range Gly^{50}-Ser^{54} shows negative $^{13}C^{\alpha}$ and $^{13}C'$ secondary shifts corresponding to 36 % of a fully formed β-strand and was defined as β in Figure 4.

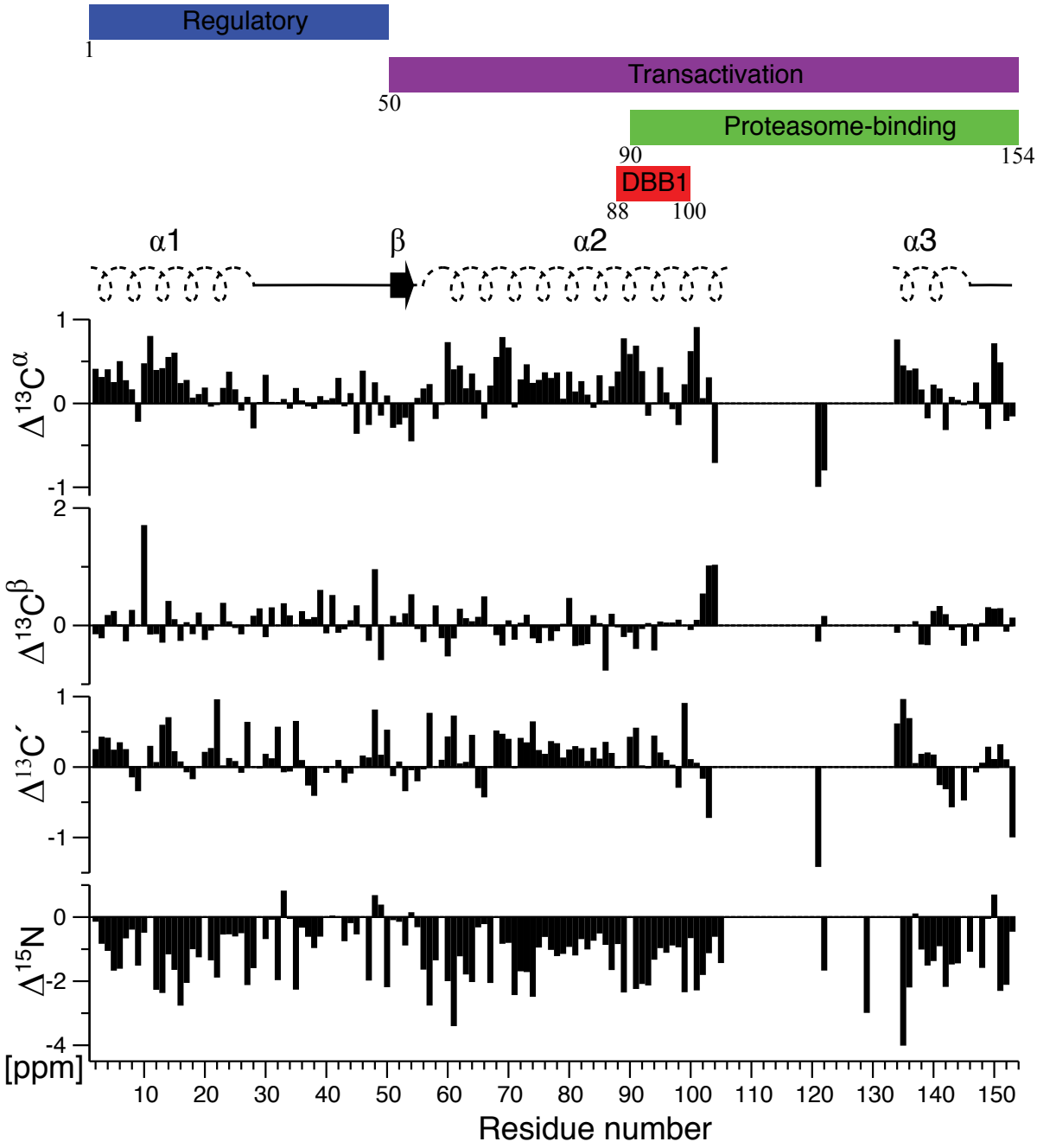

Figure 4: $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, and ^{15}N secondary shifts and derived secondary structure propensities along the amino acid sequence of HBx(9CS). A schematic representation of the functional sites of HBx is shown at the top: regulatory domain (blue) (32), transactivation domain (magenta) (11, 15, 16), proteasome-binding domain (green) (22) and the DDB1-binding domain (red) (26). Note that the assignment of residue Glu^{122} is tentative.

Backbone dynamics

To characterize the backbone dynamics of HBx(9CS), we used ^{15}N R_1, R_2 relaxation rates and $\{^1H\}$-^{15}N NOEs measured on a 600 and 900 MHz spectrometer (Figure 5). Evaluation of these data revealed several prominent details.

The R_1 values are largely independent of the magnetic field strength. This indicates that the residues move faster than the rotational correlation time of the R_1 maximum, i.e. about 2-3 ns. Furthermore the R_1 values show a significant variability within the amino acid sequence. The largest values are found in the α1, β, α2 and α3 regions indicating the slowest motions. In contrast, R_1 values are smaller in the amino acid region connecting α1 and β, as well as in the C-terminal region after α3. These faster motions are corroborated by respective reductions in the R_2 and $\{^1H\}$-^{15}N NOE values. The unfolded state and its overall very fast motions are also evident from all $\{^1H\}$-^{15}N NOE values, which are significantly lower than the values of > ~0.8 observed in folded proteins (33). The field dependence of the $\{^1H\}$-^{15}N NOEs is as expected for such fast motions with values in the range of -0.5 to 0 at 600 MHz, but between 0 and +0.5 at 900 MHz.

The behavior of the R_2 values reveals additional slower motions. This is visible from increased R_2 values at higher fields in the region of residues Ala^{85}-Ser^{104} at the end of α2 and Gly^{135}-Arg^{138} at the beginning of α3. This behavior is a clear sign of broadening due to conformational exchange on the microsecond time scale of chemical shifts. This is also consistent with the absence of observable resonances between α2 and α3, which are apparently broadened beyond detection by this exchange.

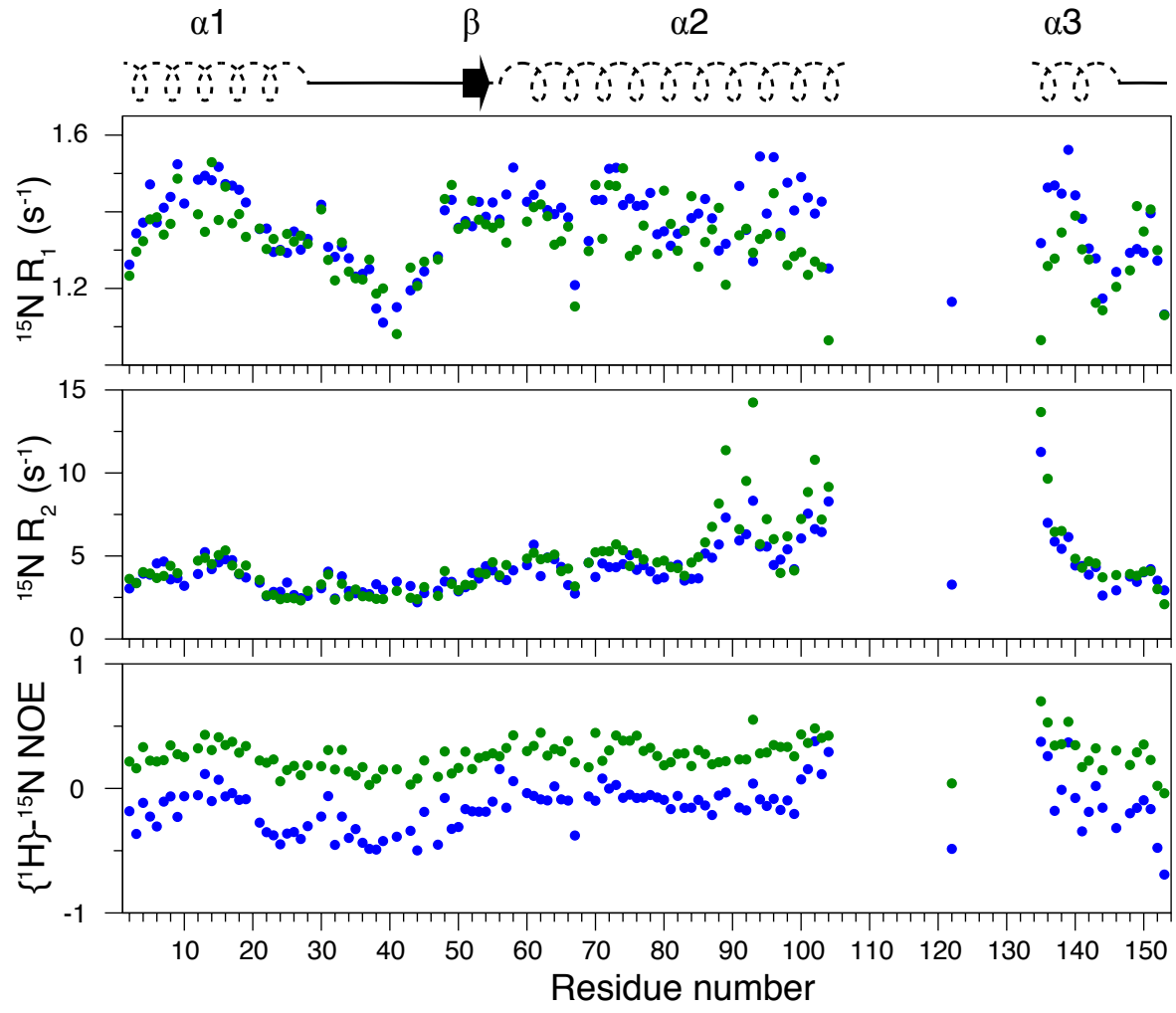

Figure 5: ^{15}N relaxation data of HBx(9CS). ^{15}N R_1, R_2 and {^{1}H}-^{15}N NOE values are shown versus the primary sequence. Measurements were carried out at 600 MHz (blue) and 900 MHz (green). Secondary structure propensities derived from chemical shifts (Figure 4) are shown at the top. Note that the assignment of residue Glu^{122} is tentative.

Disorder prediction from primary sequence in comparison to experimental data

Disorder predictors based on the amino acid sequence are often used to identify potential disordered regions in proteins. Three of these predictors were tested on the sequence of HBx(9CS):

GlobPlot2 (34) determines the tendency for disorder in a protein based on the cumulative sum of the propensity for coil conformation of each amino acid in the protein sequence. Potential disordered (ordered) regions are identified from sequence regions with increasing (decreasing) coil propensities (see Figure 6B).

IUPred (35) determines the probability of a given amino acid to be disordered from the number of stabilizing interactions with its sequential neighbors. These are estimated based on the amino acid pairing frequencies in known crystal structures. Disordered (ordered) regions are then identified from contiguous disorder probabilities higher (lower) than 0.5 (see Figure 6C).

PONDR VLXT (36, 37) uses amino acid properties such as hydropathy or net charge to calculate a disorder score. Disorder (order) is then identified by a score higher (lower) than 0.5 (see Figure 6D).

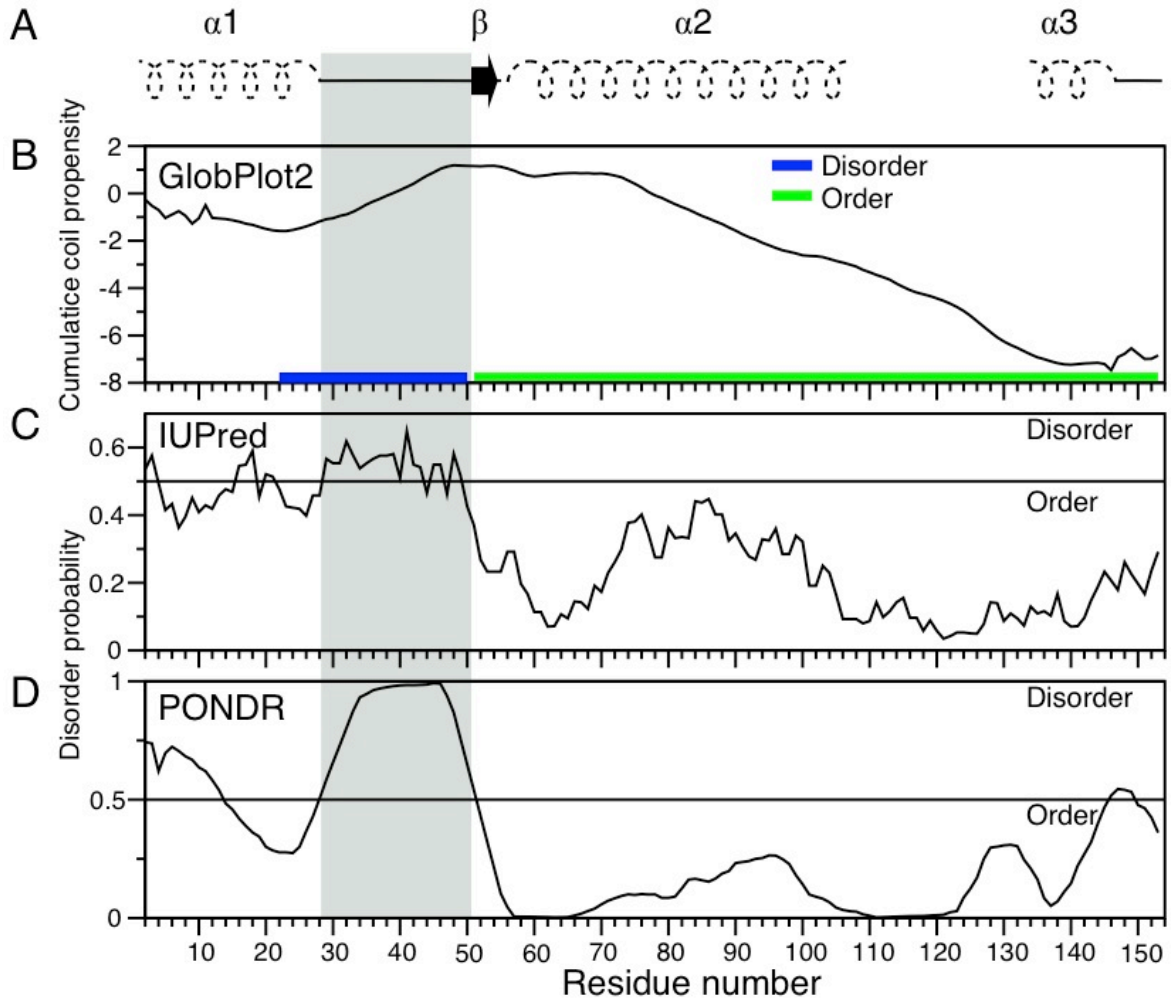

Figure 6: Prediction of structural disorder based on the HBx(9CS) primary sequence. **A** Derived secondary structure propensities based on the secondary chemical shifts. **B-D** predicted disorder propensity using: **B** GlobPlot2 **C** IUPred and **D** PONDR VLXT algorithms. The grey box highlights the coil region Arg^{28}-Gly^{50} derived from the chemical shifts.

Analysis of the HBx(9CS) primary sequence using the three disorder predictors revealed the following:

GlobPlot2. The amino acid region Gly^{22}-Gly^{50} is disordered and Ala^{51}-Ser^{153} is ordered.

IUPred. The regions Arg^{4}-Val^{15}, Gly^{22}-Arg^{28} and Gly^{50}-Ala^{154} are ordered. However, Pro^{29}-Asp^{48} is disordered.

PONDR. The ranges Ala^{2}-Arg^{13} and Arg^{28}-Ala^{51} are disordered; and the ranges Asp^{14}-Gly^{27} and Leu^{53}-Ala^{154} are ordered. The range Ala^{146}-Asn^{149} is too short to be considered disordered.

All these predictors agree on the range Arg^{28}-Gly^{50} being disordered and Leu^{53}-Ser^{153} being ordered, whereas these algorithms diverge in the range Ala^{2}-Gly^{27}. The consensus regions from the predictions agree well with the results from the NMR analysis based on chemical shifts and ^{15}N relaxation data, which indicate that the range Arg^{28}-Gly^{50} is a highly flexible coil, whereas the range Leu^{53}-Ser^{153} is more ordered with propensities for slowly moving, nascent β, α2 and α3 secondary structure elements.

Conclusions

The present NMR analysis provides the first detailed structural and dynamical characterization of HBx in solution. It is based on the nearly complete backbone and side-chain assignments, which allowed the analysis of the secondary structure propensities on the basis of chemical shifts and ^{15}N relaxation data.

HBx(9CS) is a largely unfolded protein. However, deviations from random coil chemical shifts clearly indicate that the protein contains four segments, which form transiently stable secondary structure elements ($\alpha1$, β, $\alpha2$, $\alpha3$). The ^{15}N relaxation data show the different flexibility of these segments in the nanosecond and microsecond regime. Particularly slow microsecond motions are evident from chemical exchange broadening in the region of the nascent secondary structure elements $\alpha2$ and $\alpha3$. The residual structure detected in the HBx polypeptide chain coincides with structure predictions from three widely used sequence analysis algorithms.

Interestingly, the regions with higher secondary structure propensity overlap with the functional sites of HBx. This suggests that these transiently stable structural elements represent the seeds for recognition of a partner and the subsequent transition to a fully folded conformation by conformational selection (38).

Materials and methods

HBx(9CS) construct

The full-length mutant HBx(9CS) was constructed by site-directed mutagenesis as an MGSS-HHHHHH-SSGR-ENLYFQ-SGS-HBx(-Cys)(Ala^{2}-Ala^{154}) fusion protein. As compared to the wild type (green), all cysteines were replaced by serines and an N-terminal hexahistidine-tag (red) together with a TEV cleavage site (blue) were introduced. The final sequence is:

MGSSHHHHHHSSGRENLYFQGSGSAARLSSQLD-
PARDVLSLRPVGAESRGRPFSGSLGTLSSPSPSAVPTDHGAHLSLRGLPVCAFSSAG-
PCALRFTSARRMETTVNAHQILPKVLHKRTLGLSAM-
STTDLEAYFKDCLFKDWEELGEEIRLKVFVLGGCRHKLVSAPAPSNFFTSA

Expression and purification

The full-length mutant HBx(9CS) was overexpressed in *Escherichia coli*. The protein in form of inclusion bodies was resuspended in denaturing buffer (10 mM Tris, 8 M urea, 10 mM EDTA, 2.8 mM FOS-12 (Anatrace), 100 mM DHPC (Avanti Polar Lipids, Inc.), 10 mM DDT, pH 8.0) and sonicated on ice for 5 min (5 sec on and 10 sec off). The sample was acidified to pH 4.0 using HCl and the soluble fraction was separated by centrifugation (10'000 xg for 10 min). This was followed by desalting using a sodium acetate column (5, 10 or 25, GE Healthcare Life Sciences). The equilibration and elution buffer was 20 mM sodium acetate, 2 mM EDTA, 2.8 mM FOS-12, 100 mM DHPC, 10% D_2O, pH 4.0.

NMR experiments

NMR samples were measured in Shigemi NMR tubes: 250 µL volumes of either 589 µM or 886 µM HBx(9CS), 20 mM sodium acetate, 2 mM EDTA, 2.8 mM FOS-12, 100 mM DHPC, 10% D_2O, pH 4.0. All NMR measurements were carried out at 310 K on a 600 MHz or 900 MHz Bruker Avance III spectrometer equipped with a cryogenic probe.

Temperature calibration was performed as described using a standard methanol sample (39). ^{1}H, ^{15}N, and ^{13}C chemical shifts were referenced relative to the ^{2}H lock resonance of water at 310 K.

NMR Resonance Assignments

Backbone assignments were obtained from standard triple-resonance two-, three- and four-dimensional experiments, including HNCO, HNCA, HN(CO)CA, CBCANH, (H)NNH, CBCA(CO)NH, C(CO)NH, and H(CCO)NH (40). For details about duration and acquisition times see Table S3 (**Supplementary Information**). All NMR data were processed using the NMRPipe suite of programs (41). Spectra were displayed and analyzed with the program SPARKY (42).

In brief, the assignment strategy consists of classifying all the HN, N-frequency pairs according to cross peaks in the HNCO spectrum, which is the most well resolved and most sensitive of the above set of experiments. Corresponding C_i^{α}, C_{i-1}^{α}, C_i^{β}, C_{i-1}^{β}, C_{i-1}^{γ}, C_{i-1}^{δ}, H_{i-1}^{α}, H_{i-1}^{β}, H_{i-1}^{γ}, H_{i-1}^{δ}, and HN_i frequencies are extracted from the above-mentioned experiments. These frequencies are associated with one HNCO cross peak, assuming that it represents a unique amide bond within the protein backbone. The spin systems associated with each HNCO peak are then matched against all others for possible interresidue correlations to generate sequential stretches of amino acid spin systems. These stretches are, in turn, assigned to locations in the primary sequence of the protein based on the amino acid type probability derived from the C^{α} and C^{β} chemical shifts of these stretches (43).

NMR Relaxation Experiments and Analysis

Standard ^{15}N relaxation data were recorded as described elsewhere (40). Peak intensities were determined using the program nLinLS contained in the NMRPipe package. Decay curves were fitted by an in-house written routine implemented in MATLAB (MathWorks, Inc.) using a nonlinear search minimization (Fmin) and a Monte Carlo estimation of the errors. Experimental errors were estimated from the deviations of two repeated measurements. For details about duration and acquisition times see Table S4 and S5 (**Supplementary Information**).

References

1. Levrero M, Zucman-Rossi J (2016) Mechanisms of HBV-induced hepatocellular carcinoma. *J Hepatol* 64(1 Suppl):S84–S101.

2. Ferlay J, et al. (2014) Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* 136(5):E359–E386.

3. Chisari FV, Ferrari C, Mondelli MU (1989) Hepatitis B virus structure and biology. *Microb Pathog* 6(5):311–325.

4. Meyers ML, Trepo LV, Nath N, Sninsky JJ (1986) Hepatitis B virus polypeptide X: expression in Escherichia coli and identification of specific antibodies in sera from hepatitis B virus-infected humans. *J Virol* 57(1):101-9.

5. Hoare J, Henkler F, Dowling JJ (2001) Subcellular localisation of the X protein in HBV infected hepatocytes. *J Med Virol* 64(4):419-26.

6. Henkler F, et al. (2001) Intracellular localization of the hepatitis B virus HBx protein. *J Gen Virol* 82(Pt 4):871–882.

7. Gong D-Y, et al. (2013) Role and functional domain of hepatitis B virus X protein in regulating HBV transcription and replication in vitro and in vivo. *Viruses* 5(5):1261-1271.

8. Colgrove R, Simon G, Ganem D (1989) Transcriptional activation of homologous and heterologous genes by the hepatitis B virus X gene product in cells permissive for viral replication. *J Virol* 63(9):4019–4026.

9. Feitelson MA, Bonamassa B, Arzumanyan A (2014) The roles of hepatitis B virus-encoded X protein in virus replication and the pathogenesis of chronic liver disease. *Expert Opin Ther Targets* 18(3):293–306.

10. Barnabas S, Andrisani OM (2000) Different regions of hepatitis B virus X protein are required for enhancement of bZip-mediated transactivation versus transrepression. *J Virol* 74(1):83–90.

11. Choi BH, Park GT, Rho HM (1999) Interaction of hepatitis B viral X protein and CCAAT/ enhancer-binding protein alpha synergistically activates the hepatitis B viral enhancer II/pregenomic promoter. *J Biol Chem* 274(5):2858–2865.

12. Elmore LW, et al. (1997) Hepatitis B virus X protein and p53 tumor suppressor interactions in the modulation of apoptosis. *Proc Natl Acad Sci USA* 94(26):14707-14712.

13. Koike K, et al. (1998) Compensatory apoptosis in preneoplastic liver of a transgenic mouse model for viral hepatocarcinogenesis. *Cancer Lett* 134(2):181–186.

14. Shih WL, Kuo ML, Chuang SE, Cheng AL, Doong SL (2000) Hepatitis B virus X protein inhibits transforming growth factor-beta -induced apoptosis through the activation of phosphatidylinositol 3-kinase pathway. *J Biol Chem* 275(33):25858-25864.

15. Spandau DF, Lee CH (1988) trans-activation of viral enhancers by the hepatitis B virus X protein. *J Virol* 62(2):427–434.

16. Tang H, et al. (2005) The transcriptional transactivation function of HBx protein is important for its augmentation role in hepatitis B virus replication. *J Virol* 79(9):5548-5556.

17. Rossner MT (1992) Hepatitis B virus X-gene product: A promiscuous transcriptional activator. *J Med Virol* 36(2):101-17.

18. Lee TH, Elledge SJ, Butel JS (1995) Hepatitis B virus X protein interacts with a probable cellular DNA repair protein. *J Virol* 69(2):1107–1114.

19. Lin-Marq N, Bontron S, Leupin O, Strubin M (2001) Hepatitis B virus X protein interferes with cell viability through interaction with the p127-kDa UV-damaged DNA-binding protein. *Virology* 287(2):266–274.

20. Sitterlin D, Bergametti F, Tiollais P, TENNANT BC, Transy C (2000) Correct binding of viral X protein to UVDDB-p127 cellular protein is critical for efficient infection by hepatitis B viruses. *Oncogene* 19(38):4427–4431.

21. Iovine B, Iannella ML, Bevilacqua MA (2011) Damage-specific DNA binding protein 1 (DDB1): a protein with a wide range of functions. *Int J Biochem Cell Biol* 43(12):1664–1667.

22. Minor MM, Slagle BL (2014) Hepatitis B virus HBx protein interactions with the ubiquitin proteasome system. *Viruses* 6(11):4683–4702.

23. Zhang Z, et al. (2000) Structural and functional characterization of interaction between hepatitis B virus X protein and the proteasome complex. *J Biol Chem* 275(20):15157-15165.

24. Lee TH, Elledge SJ, Butel JS (1995) Hepatitis B virus X protein interacts with a probable cellular DNA repair protein. *J Virol* 69(2):1107–1114.

25. Keasler VV, Slagle BL (2008) The interaction of HBx with cellular DDB1. In The Pleiotropic Functions of the Viral Protein HBx in Hepatitis B Virus Infection and the Development of Liver Cancer. *J Virol* 79(7):4238-45.

26. Li T, Robert EI, van Breugel PC, Strubin M, Zheng N (2010) A promiscuous alpha-helical motif anchors viral hijackers and substrate receptors to the CUL4-DDB1 ubiquitin ligase machinery. *Nat Struct Mol Biol* 17(1):105–111.

27. Li T, Chen X, Garbutt KC, Zhou P, Zheng N (2005) Structure of DDB1 in Complex with a Paramyxovirus V Protein: Viral Hijack of a Propeller Cluster in Ubiquitin Ligase. *Cell* 124(1):105–117.

28. Rui E, de Moura PR, Kobarg J (2001) Expression of deletion mutants of the hepatitis B virus protein HBx in E. coli and characterization of their RNA binding activities. *Virus Res* 74(1-2):59–73.

29. Hildt E, Hofschneider PH, Urban S (1996) The role of hepatitis B virus (HBV) in the development of hepatocellular carcinoma. *Sem Virol* 7(5):333–347.

30. Rui E, Moura PR de, Gonçalves K de A, Kobarg J (2005) Expression and spectroscopic analysis of a mutant hepatitis B virus onco-protein HBx without cysteine residues. *J Virol Methods* 126(1-2):65–74.

31. Lee S-H, et al. (2012) Structural characterization of an intrinsically unfolded mini-HBX protein from hepatitis B virus. *Mol Cells* 34(2):165–169.

32. Murakami S, Cheong JH, KANEKO S (1994) Human hepatitis virus X gene encodes a regulatory domain that represses transactivation of X protein. *J Biol Chem* 269(21):15118–15123.

33. Kay LE, Torchia DA, Bax A (1989) Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry* 28(23):8972–8979.

34. Linding R, Russell RB, Neduva V, Gibson TJ (2003) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res* 31(13):3701–3708.

35. Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21(16):3433–3434.

36. Dunker AK, et al. (2001) Intrinsically disordered protein. *J Mol Graph Model* 19(1):26–59.

37. Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN (2010) PONDR-FIT: A meta-predictor of intrinsically disordered amino acids. *Biochim Biophys Acta* 1804(4):996–1010.

38. Dogan J, Gianni S, Jemth P (2014) The binding mechanisms of intrinsically disordered proteins. *Phys Chem Chem Phys* 16(14):6323–6331.

39. Raiford DS, Fisk CL, Becker ED (1979) Calibration of methanol and ethylene glycol nuclear magnetic resonance thermometers. *Anal Chem* 51(12):2050–2051.

40. Grzesiek S, et al. (1997) Refined solution structure and backbone dynamics of HIV 1 Nef. *Protein Sci* 6(6):1248–1263.

41. Delaglio F, et al. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* 6(3):277–293.

42. Goddard T, Kneller D (2008) *SPARKY 3.*

43. Grzesiek S, Bax A (1993) Amino acid type determination in the sequential assignment procedure of uniformly 13C/15N-enriched proteins. *J Biomol NMR* 3:185-204.

Structural and dynamic characterization of the hepatitis B virus X protein (HBx)

Supplementary Information

Leonildo Delgado[1]*, Fabio Casagrande*[2]*, Cristoph Bieniossek*[2] *and Stephan Grzesiek*[1]

[1]Focal Area Structural Biology and Biophysics, Biozentrum, University of Basel, CH-4056 Basel, Switzerland

[2]Roche Innovation Center, F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, CH-4070 Basel, Switzerland

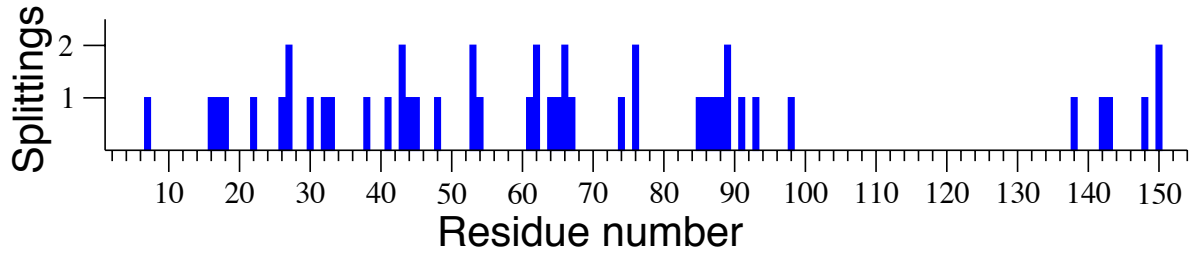

Figure S1: Number of splitted peaks of ^{1}H-^{15}N amide resonances attributed to proline cis-trans isomerization equilibria.

Table S1: Sequence specific assignment of $^{1}H^{N}$, ^{15}N, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, and ^{13}CO chemical shifts of HBx(9CS).

Residue	$^{1}H^{N}$ (ppm)	^{15}N (ppm)	$^{13}C^{\alpha}$ (ppm)	$^{13}C^{\beta}$ (ppm)	^{13}CO (ppm)
A2	8.284	125.506	53.008	19.032	177.786
A3	8.052	122.289	52.934	19.042	178.033
R4	8.078	119.379	56.367	30.723	176.641
L5	8.171	122.461	55.581	42.417	177.69
S6	8.229	115.634	58.953	63.752	174.975
S7	8.179	117.004	58.893	63.742	174.692
Q8	8.203	121.211	56.144	29.423	175.77
L9	8.018	121.75	55.129	42.544	176.534
D10	8.222	121.047	51.885	40.323	
P11			63.963	32.086	177.06
A12	8.204	121.696	52.976	18.949	178.067
R13	7.92	118.229	56.512	30.844	176.339
D14	8.229	119.882	54.142	39.867	176.086
V15	7.956	119.436	63.037	32.409	176.324
L16	8.122	123.279	55.624	42.111	177.4
S17	8.017	115.276	58.464	63.764	174.313
L18	8.017	123.179	55.131	42.43	176.849
R19	8.086	121.879	53.933	30.395	
P20			63.055	31.982	176.965
V21	8.209	120.22	62.585	32.816	176.875
G22	8.444	112.22	45.336		174.163
A23	8.174	123.761	52.838	19.338	178.043
E24	8.435	119.055	56.49	29.323	176.493
S25	8.221	116.437	58.491	63.823	174.671
R26	8.266	122.527	56.332	30.715	176.684
G27	8.329	109.234	45.185		173.646
R28	8.086	121.192	53.957	30.324	
P29	8.299		63.26	31.959	176.586
F30	8.298	119.979	57.9	39.456	175.859
S31	8.178	117.526	58.367	63.989	174.724
G32	7.829	110.123	45.451		174.039
S33	8.16	115.416	58.407	64.005	174.733

L34	8.318	123.655	55.573	42.392	177.841
G35	8.314	108.744	45.546		174.387
T36	7.957	113.507	61.949	69.85	174.709
L37	8.213	124.006	55.376	42.418	177.207
S38	8.163	116.02	58.178	64.027	174.075
S39	8.184	118.463	56.373	63.584	
P40			63.168	32.115	176.753
S41	8.349	117.323	56.355	63.514	
P42			63.535	32.095	177.001
S43	8.217	115.199	58.276	63.922	174.167
A44	8.194	125.954	52.341	19.461	177.288
V45	8.007	120.459	59.875	32.613	
P46			63.34	32.171	177.14
T47	8.138	113.719	61.841	69.841	174.359
D48	8.263	121.825	53.876	40.611	176.075
H49	8.522	119.034	55.679	28.629	174.924
G50	8.436	109.027	45.392		173.76
A51	8.056	123.277	52.538	19.345	177.489
H52	8.464	117.148	55.093	28.965	174.252
L53	8.287	123.087	55.255	42.734	176.785
S54	8.301	116.74	57.856	64.151	174.369
L55	8.368	123.932	55.275	42.356	177.039
R56	8.224	120.136	56.56	30.489	176.499
G57	8.278	108.577	45.169		173.774
L58	7.889	122.112	52.977	42.171	
P59			62.949	31.52	176.917
V60	8.148	119.48	63.07	32.548	176.674
S61	8.223	117.235	58.721	63.649	174.826
A62	8.144	124.97	53.118	19.123	177.419
F63	7.953	116.88	57.797	39.465	175.827
S64	7.996	115.658	58.569	64.048	174.656
S65	8.254	117.456	58.612	63.868	174.151
A66	8.16	125.024	52.53	19.678	177.648
G67	8.094	107.523	44.89		
P68			63.797	32.065	177.721
S69	8.345	115.411	59.246	63.566	174.928
A70	8.191	124.909	53.128	19.228	177.891
L71	7.903	119.089	55.413	42.093	177.21
R72	7.908	120.035	56.361	30.795	176.136
F73	8.122	119.925	57.93	39.553	175.998
T74	8.002	114.243	62.125	69.861	174.621
S75	8.211	117.478	58.715	63.861	174.511
A76	8.226	125.519	52.983	19.184	177.793
R77	8.123	119.38	56.408	30.674	176.476
R78	8.222	121.747	56.424	30.715	176.439
M79	8.344	121.397	55.732	32.959	176.226

E80	8.388	121.57	56.471	29.639	176.395
T81	8.173	114.727	62.066	69.769	174.73
T82	8.111	116.529	62.126	69.836	174.65
V83	8.113	122.087	62.615	32.728	175.769
N84	8.391	122.075	53.116	38.965	175.115
A85	8.267	124.342	53.107	19.152	177.656
H86	8.392	116.195	55.524	28.74	174.339
Q87	8.19	120.522	56.061	29.702	175.657
I88	8.176	121.993	61.264	38.679	175.864
L89	8.254	126.22	53.71	41.571	
P90			63.465	32.053	176.913
K91	8.161	120.326	56.934	32.959	176.91
V92	7.989	120.215	62.511	32.705	175.791
L93	8.168	124.522	55.071	42.515	176.874
H94	8.346	118.826	55.251	29.013	174.306
K95	8.312	122.386	56.707	33.264	176.42
R96	8.436	122.298	56.368	30.957	176.28
T97	8.137	115.452	61.814	69.817	174.322
L98	8.274	124.128	55.404	42.722	177.257
G99	8.275	108.697	45.498		174.441
L100	8.168	121.316	55.971	42.482	177.716
S101	8.262	114.364	59.219	63.767	174.337
A102	8.072	124.035	52.629	19.422	177.606
M103	7.926	118.41	55.962	34.062	175.662
S104	8.529	116.87	57.725	64.887	
T105	8.191	115.255			
T106					
D107					
L108					
E109					
A110					
F111					
F112					
K113					
D114					
S115					
L116					
F117					
K118					
D119					
W120					
E121			55.154	28.885	174.176
E122	8.497	119.114	55.186	29.004	
L123					
G124					
E125					

E126					
I127					
R128					
L129	8.013	121.813			
K130					
V131					
F132					
V133					
L134			56.132	42.185	178.308
G135	8.129	107.142	45.926		175.011
G136	8.115	108.1	45.701		174.497
S137	8.158	115.165	58.858	63.758	174.856
R138	8.187	121.445	56.44	30.449	176.196
H139	8.298	118.271	55.332	29.055	174.074
K140	8.236	121.9	56.439	33.251	176.136
L141	8.304	123.531	55.35	42.462	176.99
V142	8.038	119.977	62.057	33.018	175.725
S143	8.221	118.819	58.011	64.035	173.518
A144	8.162	126.622	50.467	18.548	
P145			62.897	32.004	176.111
A146	8.262	124.911	50.385	18.434	
P147			63.396	32.033	176.995
S148	8.243	114.672	58.647	63.811	174.424
N149	8.259	119.735	53.367	38.944	174.928
F150	8.07	119.673	58.562	39.656	175.264
F151	8.041	118.728	57.844	39.59	175.564
T152	7.919	114.345	61.726	69.922	174.127
S153	8.168	118.015	58.303	64.153	173.365
A154	8.019	129.732	53.189	19.963	

Table S2: ^{15}N relaxation data recorded at 310 K on a 600 and 900 MHz Bruker Avance III spectrometer.

	900 MHz						600 MHz					
Residue	R_1 [s^{-1}]	SD* [s^{-1}]	R_2 [s^{-1}]	SD* [s^{-1}]	NOE	SD*	R_1 [s^{-1}]	SD* [s^{-1}]	R_2 [s^{-1}]	SD* [s^{-1}]	NOE	SD*
A2	1.233	0.029	3.618	0.021	0.216	0.010	1.262	0.011	3.045	0.081	-0.183	0.021
A3	1.296	0.016	3.363	0.151	0.162	0.042	1.344	0.015	3.380	0.033	-0.366	0.001
R4	1.323	0.009	4.033	0.019	0.332	0.025	1.372	0.009	3.939	0.077	-0.116	0.016
L5	1.380	0.080	3.950	0.031	0.223	0.039	1.471	0.001	3.871	0.032	-0.226	0.089
S6	1.386	0.011	3.668	0.126	0.216	0.007	1.372	0.039	4.558	0.001	-0.306	0.024
S7	1.340	0.037	3.798	0.221	0.226	0.011	1.411	0.040	4.662	0.022	-0.106	0.027
Q8	1.369	0.030	4.407	0.056	0.346	0.023	1.439	0.023	3.590	0.163	-0.065	0.021
L9	1.487	0.004	3.968	0.038	0.274	0.012	1.524	0.026	3.661	0.016	-0.228	0.043
D10					0.251	0.010	1.422	0.058	3.206	0.217	-0.063	0.011
P11												
A12	1.394	0.013	4.722	0.045	0.321	0.013	1.484	0.011	3.912	0.032	-0.055	0.000
R13	1.348	0.067	4.884	0.131	0.431	0.064	1.494	0.036	5.236	0.166	0.116	0.094
D14	1.529	0.046	4.516	0.208	0.309	0.019	1.482	0.012	4.206	0.095	-0.103	0.009
V15	1.378	0.035	5.054	0.096	0.412	0.021	1.517	0.076	4.614	0.035	0.069	0.040
L16	1.466	0.028	5.341	0.120	0.348	0.032	1.472	0.086	4.807	0.313	-0.065	0.041
S17	1.370	0.021	4.426	0.119	0.375	0.001	1.468	0.015	4.746	0.021	-0.039	0.027
L18	1.394	0.034	3.934	0.068	0.287	0.033	1.457	0.037	3.897	0.106	-0.093	0.012
R19	1.335	0.022	4.427	0.018	0.340	0.075	1.424	0.091	3.705	0.118	-0.086	0.034
P20	1.357	0.012										
V21	1.303	0.054	3.551	0.034	0.225	0.026	1.355	0.001	3.387	0.008	-0.275	0.014
G22	1.329	0.010	2.632	0.088	0.207	0.052	1.356	0.028	2.578	0.087	-0.352	0.045
A23	1.299	0.003	2.664	0.063	0.232	0.056	1.295	0.042	2.841	0.011	-0.377	0.013
E24	1.342	0.006	2.403	0.089	0.057	0.035	1.301	0.004	2.845	0.042	-0.449	0.003
S25	1.323	0.018	2.466	0.062	0.147	0.017	1.293	0.004	3.404	0.002	-0.363	0.071
R26			2.441	0.179	0.181	0.022	1.349	0.021	2.644	0.048	-0.350	0.008
G27	1.338	0.075	2.326	0.113	0.106	0.067	1.301	0.040	2.453	0.013	-0.405	0.036
R28	1.317	0.021	2.899	0.138	0.185	0.016	1.329	0.015	2.597	0.054	-0.303	0.057
P29												
F30	1.406	0.025	3.289	0.065	0.178	0.044	1.419	0.006	3.060	0.074	-0.227	0.027
S31	1.274	0.010	3.883	0.003	0.308	0.004	1.308	0.036	4.061	0.137	-0.062	0.075
G32	1.221	0.001	2.364	0.047	0.152	0.003	1.282	0.022	2.437	0.084	-0.454	0.058
S33	1.320	0.028	3.329	0.080	0.310	0.003	1.309	0.013	3.780	0.042	-0.225	0.018
L34	1.244	0.008	2.567	0.053	0.136	0.013	1.279	0.006	2.893	0.088	-0.397	0.070
G35	1.226	0.027	2.984	0.007	0.104	0.015	1.231	0.050	2.749	0.026	-0.326	0.007
T36	1.223	0.057	2.575	0.013	0.171	0.041	1.238	0.021	2.811	0.141	-0.436	0.068
L37	1.275	0.020	2.557	0.070	0.028	0.038	1.250	0.017	2.702	0.026	-0.485	0.043
S38	1.186	0.001	2.425	0.015	0.077	0.018	1.148	0.041	3.294	0.045	-0.491	0.066
S39	1.200	0.026	2.413	0.069	0.152	0.048	1.111	0.047	2.961	0.016	-0.423	0.026
P40												
S41	1.081	0.002	2.900	0.046	0.153	0.008	1.151	0.017	3.453	0.008	-0.388	0.012
P42												

S43	1.254	0.003	2.480	0.084	0.031	0.037	1.195	0.029	3.190	0.000	-0.341	0.006
A44	1.207	0.009	2.402	0.004	0.080	0.095	1.215	0.003	2.204	0.007	-0.498	0.044
V45	1.270	0.008	3.133	0.082	0.225	0.029	1.244	0.005	2.757	0.044	-0.188	0.081
P46												
T47	1.276	0.057	2.592	0.155	0.093	0.034	1.284	0.053	2.907	0.163	-0.453	0.041
D48	1.434	0.006	4.089	0.021	0.297	0.011	1.404	0.016	3.461	0.116	-0.076	0.000
H49	1.470	0.019	3.316	0.279	0.120	0.010	1.431	0.048	3.444	0.045	-0.325	0.053
G50	1.355	0.037	2.939	0.072	0.164	0.012	1.359	0.107	2.873	0.018	-0.310	0.024
A51	1.368	0.081	3.259	0.032	0.296	0.072	1.376	0.003	3.122	0.062	-0.166	0.009
H52	1.429	0.102	3.245	0.004	0.155	0.065	1.362	0.030	3.975	0.002	-0.185	0.021
L53	1.380	0.024	3.993	0.019	0.245	0.018	1.426	0.082	3.641	0.063	-0.187	0.012
S54	1.367	0.026	3.909	0.133	0.259	0.050	1.387	0.035	4.383	0.138	-0.187	0.016
L55	1.359	0.004	4.624	0.153	0.281	0.035	1.424	0.020	4.131	0.128	-0.106	0.004
R56	1.370	0.018	3.833	0.047	0.258	0.018	1.380	0.005	3.723	0.076	0.155	0.050
G57	1.319	0.022	4.454	0.045	0.325	0.031	1.445	0.018	3.544	0.185	-0.155	0.016
L58					0.426	0.022	1.516	0.069	4.137	0.051	0.059	0.125
P59												
V60	1.375	0.003	4.842	0.049	0.299	0.000	1.426	0.013	4.446	0.142	-0.038	0.043
S61	1.412	0.028	5.189	0.322	0.342	0.003	1.444	0.037	5.681	0.176	-0.060	0.030
A62	1.419	0.016	4.800	0.045	0.447	0.045	1.471	0.017	3.784	0.034	-0.088	0.047
F63	1.389	0.009	4.881	0.159	0.263	0.037	1.405	0.004	4.906	0.067	-0.096	0.119
S64	1.314	0.021	5.088	0.223	0.316	0.011	1.394	0.001	4.807	0.138	0.016	0.022
S65	1.323	0.006	4.084	0.189	0.298	0.036	1.411	0.034	4.338	0.135	-0.088	0.020
A66	1.361	0.016	4.242	0.054	0.381	0.047	1.385	0.007	3.238	0.226	-0.098	0.047
G67	1.153	0.031	3.176	0.251	0.209	0.062	1.209	0.044	2.730	0.000	-0.378	0.064
P68												
S69	1.297	0.031	4.589	0.117	0.170	0.020	1.324	0.023	4.595	0.050	-0.064	0.050
A70	1.470	0.065	5.217	0.081	0.446	0.052	1.431	0.036	3.741	0.142	-0.099	0.062
L71	1.330	0.054	5.298	0.271	0.222	0.009	1.431	0.078	4.549	0.109	0.079	0.014
R72	1.470	0.070	5.287	0.054	0.306	0.006	1.512	0.030	4.330	0.174	-0.002	0.013
F73	1.468	0.029	5.695	0.211	0.424	0.028	1.515	0.005	4.322	0.143	0.028	0.005
T74	1.514	0.067	5.353	0.148	0.384	0.068	1.418	0.076	4.506	0.348	-0.076	0.110
S75	1.285	0.024	4.390	0.019	0.384	0.001	1.434	0.030	5.042	0.222	-0.050	0.055
A76	1.301	0.079	5.162	0.324	0.424	0.014	1.416	0.003	4.164	0.036	-0.077	0.003
R77	1.364	0.005	4.797	0.040	0.301	0.003	1.418	0.035	4.445	0.009	-0.074	0.026
R78					0.325	0.022	1.449	0.007	4.070	0.020	-0.053	0.002
M79	1.290	0.004	4.612	0.270	0.260	0.079	1.341	0.104	3.587	0.155	-0.073	0.002
E80	1.455	0.050	4.718	0.182	0.186	0.010	1.349	0.141	3.707	0.046	-0.092	0.005
T81	1.369	0.049	4.316	0.057	0.211	0.073	1.312	0.012	4.349	0.068	-0.166	0.006
T82	1.299	0.018	4.267	0.144	0.277	0.004	1.343	0.104	4.470	0.045	-0.061	0.006
V83	1.351	0.051	3.804	0.163	0.281	0.073	1.352	0.022	3.514	0.036	-0.156	0.033
N84	1.441	0.025	4.597	0.041	0.179	0.005	1.384	0.036	3.615	0.152	-0.155	0.054
A85	1.257	0.065	4.937	0.199	0.308	0.030	1.395	0.077	3.652	0.140	-0.092	0.031
H86	1.321	0.010	5.814	0.134	0.275	0.003	1.434	0.048	5.148	0.392	-0.137	0.068
Q87	1.354	0.030	6.753	0.193	0.194	0.024	1.383	0.093	4.894	0.287	-0.214	0.003
I88	1.410	0.141	8.162	0.090	0.210	0.032	1.299	0.147	5.686	0.306	-0.058	0.022

L89	1.210	0.050	11.372	0.160	0.219	0.099	1.316	0.083	7.305	0.449	-0.033	0.116
P90												
K91	1.339	0.082	6.615	0.020	0.232	0.043	1.467	0.206	5.928	0.026	-0.154	0.012
V92	1.356	0.034	9.523	0.195	0.233	0.027	1.353	0.002	6.306	0.600	-0.177	0.084
L93	1.293	0.191	14.247	2.206	0.551	0.146	1.271	0.025	8.329	0.157	0.039	0.222
H94	1.329	0.012	5.715	0.177	0.281	0.035	1.544	0.052	5.564	0.414	-0.088	0.025
K95	1.342	0.017	7.216	0.176	0.288	0.080	1.395	0.064	5.562	0.051	-0.141	0.028
R96	1.448	0.082	6.019	0.017	0.349	0.062	1.543	0.009	4.456	0.091	-0.084	0.010
T97	1.338	0.014	3.974	0.045	0.333	0.034	1.346	0.003	4.785	0.044	-0.173	0.035
L98	1.261	0.022	6.178	0.408	0.333	0.044	1.476	0.018	5.393	0.081	-0.096	0.013
G99	1.284	0.007	4.118	0.075	0.257	0.004	1.404	0.003	4.208	0.039	-0.206	0.013
L100	1.295	0.005	7.226	0.223	0.434	0.014	1.490	0.039	6.061	0.044	0.072	0.023
S101	1.236	0.041	8.847	0.013	0.365	0.014	1.437	0.075	7.554	0.678	0.155	0.065
A102	1.270	0.092	10.788	0.012	0.483	0.001	1.395	0.027	6.625	0.131	0.380	0.157
M103	1.256	0.049	7.202	0.198	0.406	0.046	1.427	0.091	6.442	0.082	0.114	0.077
S104	1.065	0.033	9.162	0.404	0.424	0.181	1.252	0.074	8.285	0.827	0.292	0.114
T105												
T106												
D107												
L108												
E109												
A110												
F111												
F112												
K113												
D114												
S115												
L116												
F117												
K118												
D119												
W120												
E121												
E122					0.041	0.032	1.165	0.142	3.272	0.480	-0.485	0.009
L123												
G124												
E125												
E126												
I127												
R128												
L129												
K130												
V131												
F132												
V133												
L134												

G135	1.065	0.086	13.665	1.455	0.700	0.114	1.318	0.000	11.262	0.725	0.375	0.235
G136	1.258	0.021	9.656	0.044	0.530	0.129	1.463	0.024	6.996	0.689	0.259	0.028
S137	1.278	0.043	6.451	0.027	0.345	0.011	1.469	0.038	5.875	0.160	-0.180	0.007
R138	1.346	0.008	6.501	0.036	0.355	0.013	1.447	0.013	5.431	0.073	-0.011	0.015
H139					0.535	0.002	1.561	0.020	6.138	0.062	0.369	0.002
K140	1.390	0.011	4.834	0.019	0.348	0.014	1.443	0.030	4.421	0.125	-0.077	0.005
L141	1.302	0.053	4.295	0.014	0.172	0.012	1.382	0.040	4.395	0.140	-0.345	0.101
V142	1.276	0.019	4.671	0.266	0.222	0.028	1.304	0.022	3.870	0.227	-0.188	0.086
S143	1.163	0.054	4.551	0.173	0.321	0.051	1.279	0.054	4.331	0.076	0.019	0.062
A144	1.143	0.022	3.714	0.096	0.147	0.021	1.174	0.028	2.619	0.268	-0.155	0.157
P145												
A146	1.204	0.020	3.845	0.112	0.304	0.052	1.243	0.037	2.931	0.214	-0.318	0.113
P147												
S148	1.247	0.053	3.909	0.064	0.188	0.001	1.293	0.080	3.757	0.004	-0.199	0.112
N149	1.415	0.020	3.810	0.159	0.290	0.036	1.303	0.020	3.443	0.018	-0.157	0.025
F150	1.349	0.014	4.061	0.186	0.352	0.005	1.293	0.020	4.017	0.009	-0.095	0.007
F151	1.407	0.035	4.090	0.318	0.229	0.064	1.396	0.014	4.212	0.108	-0.167	0.059
T152	1.300	0.017	3.001	0.095	0.020	0.018	1.273	0.016	3.509	0.024	-0.478	0.161
S153	1.130	0.005	2.099	0.058	-0.040	0.023	1.132	0.005	2.936	0.059	-0.691	0.021
A154												

* The experimental errors were estimated as the root mean square deviation of two experiments.

Table S3: NMR double and triple-resonance experiments duration and acquisition times for the assignment of HBx(9CS)

Experiment	Time (h)	Acquisition times (ms)			
HSQC	0.3	90 (N)	107.5 (HN)		
HNCO	11	27 (N)	79 (HN)	49.6 (CO)	
HNCA	39.83	22.3 (N)	79.9 (HN)	9 (CA)	
HN(CO)CA	24.65	28.6 (N)	79.9 (HN)	9.1 CA)	
CBCANH	17.47	21.8 (N)	80 (HN)	6.6 (CBCA)	
CBCA(CO)NH	10.93	27 (N)	80 (HN)	6.6 (CBCA)	
C(CO)NH	42.25	25.85 (N)	79.9 (HN)	6.6 (C)	
H(CCO)NH	39	25.85 (N)	80 (HN)	15 (H)	
(H)NNH	67.65	26.4 (N1)	26.4 (N2)	25 (HN1)	82.9 (HN2)

Table S4: ^{15}N relaxation experiment duration and acquisition times recorded on the 600 MHz Bruker Avance III spectrometer.

Experiment	Time (h)	Acquisition times (ms)	
^{15}N T_1	24.4	107 (HN)	99 (N)
^{15}N T_2	24.5	107 (HN)	99 (N)
{^{1}H}-^{15}N NOE	44.3	107 (HN)	66 (N)

Table S4: ^{15}N relaxation experiment duration and acquisition times recorded on the 900 MHz Bruker Avance III spectrometer.

Experiment	Time (h)	Acquisition times (ms)	
^{15}N T_1	53.6	189 (HN)	110 (N)
^{15}N T_2	53.9	189 (HN)	110 (N)
{^{1}H}-^{15}N NOE	11.9	107 (HN)	44 (N)

4 Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS

Original publication

Aznauryan M, Delgado L, Soranno A, Nettels D, Huang J, Labhardt AM, Grzesiek S and Schuler B

Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS

Proc Natl Acad Sci USA 113(37):E5389–98 (2016)

Introduction

To fulfill their function, proteins have to be assembled into sophisticated three-dimensional structures. As shown by Anfinsen four decades ago (1), the structure of a protein is dictated by its amino acid sequence, since the protein unfolded in a denaturant regains its three-dimensional structure when brought into physiological buffer conditions. However, a complete understanding of this protein folding process is still missing, despite strong progress in recent years using different experimental and computational techniques (2-4).

A prerequisite for understanding protein folding is the accurate structural and dynamical description of the denatured state, the partially folded intermediates and of course also the native state. However, a precise description of the unfolded state remains challenging, since it consists of an ensemble of conformations and only a limited number of measurable parameters are available to describe the very large structural space (5).

Denatured states of proteins can be created by various methods including temperature (6) or pressure (7) changes, addition of denaturants such as guanidinium chloride (8), urea (9) or acid (7), and usage of destabilizing mutations (10).

Among other biophysical methods, three experimental methods have provided the most detailed insights of the unfolded state: nuclear magnetic resonance (NMR) (5, 11), small-angle X-ray scattering (SAXS) (12) and single-molecule Förster Resonance Energy Transfer (smFRET) (13). The large number of experimental data from NMR on unfolded polypeptides provide powerful input for computer models to create representative ensembles of the unfolded state. Recently, Huang et al. used restrained ensemble structure calculations to create a minimal-size ensemble of urea-denatured ubiquitin that satisfied 419 residual dipolar couplings (RDCs) and 253 paramagnetic relaxation enhancements (PREs) restraints obtained by NMR (14).

In the current study, we have extended the calculations of structural ensembles by including restraints from the SAXS profiles (15) and 71 backbone $^{3}J_{HNH\alpha}$-couplings (16) of urea-denatured ubiquitin. The resulting ensembles are very similar to the previously calculated ones (14), as observed from the C^{α}-C^{α} contact maps that show a surprisingly similar intramolecular interactions corresponding to both native and nonnative structure elements.

To validate this model description by an independent method, we probed the distances between different segments of the polypeptide chain by smFRET. For this, we created seven double-cysteine variants of ubiquitin pairwise separated by 26 to 68 amino acids, which were labeled by fluorophores. We then used single-molecule fluorescence spectroscopy to obtain transfer efficiency histograms of ubiquitin at different urea concentrations. With increasing concentrations of urea, we observe a gradual change in the relative populations of folded and unfolded states and disappearance of the folded peak at urea concentration of ~5.5 M. How-

ever, whereas the position of the folded state signal is unaffected by the addition of urea, the unfolded peak shifts to lower transfer efficiencies, indicating a continuous expansion of the chain with increasing concentration of urea.

Remarkably, both the averages as well as the distributions of the intramolecular distances derived from smFRET and from structural ensembles calculated using the program X-PLOR-NIH based on NMR and SAXS restraints agree exceedingly well.

Single-molecule fluorescence spectroscopy additionally allowed obtaining information about the reconfiguration dynamics of unfolded ubiquitin, which is difficult to assess by NMR. To this end, we used nanosecond-fluorescence correlation spectroscopy (nsFCS) to determine the distance dynamics between the donor and acceptor fluorophores and found the reconfiguration times for all variants of ubiquitin to be in the range of 50 to 90 ns.

Overall, in this chapter we show for the first time, the combined usage of NMR, SAXS and smFRET, for the investigation of an unfolded state to quantitatively test the consistency and complementarity of the data. This provided a very comprehensive view of the unfolded-state ensemble of the protein ubiquitin.

References

1. Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* 181(4096):223–230.

2. Yu H, et al. (2012) Energy landscape analysis of native folding of the prion protein yields the diffusion constant, transition path time, and rates. *Proc Natl Acad Sci USA* 109(36):14452–14457.

3. Chung HS, McHale K, Louis JM, Eaton WA (2012) Single-molecule fluorescence experiments determine protein folding transition path times. *Science* 335(6071):981-984.

4. Liu F, Nakaema M, Gruebele M (2009) The transition state transit time of WW domain folding is controlled by energy landscape roughness. *J Chem Phys* 131(19):195101.

5. Meier S, Blackledge M, Grzesiek S (2008) Conformational distributions of unfolded polypeptides from novel NMR techniques. *J Chem Phys* 128(5):052204.

6. Dill KA, Shortle D (1991) Denatured states of proteins. *Annu Rev Biochem* 60:795-825.

7. Kamatari YO, Kitahara R, Yamada H, Yokoyama S, Akasaka K (2004) High-pressure NMR spectroscopy for characterizing folding intermediates and denatured states of proteins. *Methods* 34(1):133–143.

8. Shortle D (1996) The denatured state (the other half of the folding equation) and its role in protein stability. *FASEB J* 10(1):27-34.

9. Shortle D, Ackerman MS (2001) Persistence of native-like topology in a denatured protein in 8 M urea. *Science* 293(5529):487–489.

10. Shortle D, Chan HS, Dill KA (1992) Modeling the effects of mutations on the denatured states of proteins. *Protein Sci* 1(2):201–215.

11. Mittag T, Forman-Kay JD (2007) Atomic-level characterization of disordered protein ensembles. *Curr Opin Struct Biol* 17(1):3–14.

12. Bernadó P, Svergun DI (2012) Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Mol Biosyst* 8(1):151–167.

13. Brucale M, Schuler B, Samorì B (2014) Single-molecule studies of intrinsically disordered proteins. *Chem Rev*114(6):3281–3317.

14. Huang J-R, Grzesiek S (2010) Ensemble calculations of unstructured proteins constrained by RDC and PRE data: a case study of urea-denatured ubiquitin. *J Am Chem Soc* 132(2):694–705.

15. Gabel F, Jensen MR, Zaccaï G, Blackledge M (2009) Quantitative modelfree analysis of urea binding to unfolded ubiquitin using a combination of small angle X-ray and neutron scattering. *J Am Chem Soc* 131(25):8769–8771.

16. Vajpai N, Gentner M, Huang J-R, Blackledge M, Grzesiek S (2010) Side-chain chi(1) conformations in urea-denatured ubiquitin and protein G from (3)J coupling constants and residual dipolar couplings. *J Am Chem Soc* 132(9):3196–3203.

Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS

Mikayel Aznauryan[a,1,2], Leonildo Delgado[b,1], Andrea Soranno[a], Daniel Nettels[a], Jie-rong Huang[c], Alexander M. Labhardt[b], Stephan Grzesiek[b,3], and Benjamin Schuler[a,d,3]

[a]Department of Biochemistry, University of Zurich, 8057 Zurich, Switzerland; [b]Biozentrum, University of Basel, 4056 Basel, Switzerland; [c]Institute of Biochemistry and Molecular Biology, National Yang Ming University, Taipei City 112, Taiwan; and [d]Department of Physics, University of Zurich, 8057 Zurich, Switzerland

Edited by Attila Szabo, National Institutes of Health, Bethesda, MD, and approved July 20, 2016 (received for review May 6, 2016)

The properties of unfolded proteins are essential both for the mechanisms of protein folding and for the function of the large group of intrinsically disordered proteins. However, the detailed structural and dynamical characterization of these highly dynamic and conformationally heterogeneous ensembles has remained challenging. Here we combine and compare three of the leading techniques for the investigation of unfolded proteins, NMR spectroscopy (NMR), small-angle X-ray scattering (SAXS), and single-molecule Förster resonance energy transfer (FRET), with the goal of quantitatively testing their consistency and complementarity and for obtaining a comprehensive view of the unfolded-state ensemble. Using unfolded ubiquitin as a test case, we find that its average dimensions derived from FRET and from structural ensembles calculated using the program X-PLOR-NIH based on NMR and SAXS restraints agree remarkably well; even the shapes of the underlying intramolecular distance distributions are in good agreement, attesting to the reliability of the approaches. The NMR-based results provide a highly sensitive way of quantifying residual structure in the unfolded state. FRET-based nanosecond fluorescence correlation spectroscopy allows long-range distances and chain dynamics to be probed in a time range inaccessible by NMR. The combined techniques thus provide a way of optimally using the complementarity of the available methods for a quantitative structural and dynamical description of unfolded proteins both at the global and the local level.

protein folding | unfolded protein ensemble | Förster resonance energy transfer | nuclear magnetic resonance | small-angle X-ray scattering

Proteins exist as ensembles of interconverting conformations. Obviously, unfolded polypeptide chains, such as chemically or physically denatured proteins and intrinsically disordered proteins (IDPs), can access extremely large numbers of conformations (1). A comprehensive description of their structural and dynamical behavior is a prerequisite for understanding protein folding (2–4) and the function of IDPs in health and disease (5, 6). Due to the very large number of conformational degrees of freedom of the unfolded polypeptide ensemble, it is of utmost importance to obtain as many independent experimental parameters as possible for a quantitative description of its behavior. Three experimental methods have been particularly informative in this respect: NMR spectroscopy (2, 5, 7–9), small angle X-ray scattering (SAXS) (10, 11), and single-molecule Förster resonance energy transfer spectroscopy (single-molecule FRET) (12, 13).

NMR in solution provides very rich local structural and dynamical information at virtually any atom site with the exception of oxygen. Distance and angular information can be obtained from nuclear Overhauser enhancements (14), three-bond scalar couplings (15), paramagnetic relaxation enhancements (PREs) (16), pseudo contact shifts (17), residual dipolar couplings (RDCs) (8, 18), chemical shifts (19), and hydrogen bond scalar couplings (20). Thus, on the order of 10 geometric parameters per amino acid can be obtained with relative ease on unfolded polypeptides (21). Information on dynamics is also available from solution NMR but is restricted to three timescales: the pico- to low nanosecond range of the Larmor frequencies, the micro- to low millisecond range of the chemical shift (19), and the larger-than-seconds range of real-time NMR (22). The large number of experimental data from NMR on the geometry of unfolded polypeptide ensembles together with the knowledge on the covalent structure of the polypeptide chain provide powerful input for computer models of representative ensembles of the unfolded state. Typically, ensembles containing tens of thousands of conformers are simulated in an unrestrained manner and reduced by various methods (23–26) to smaller-size ensembles that satisfy the measured parameters. As an alternative, we have recently used restrained ensemble structure calculations (27) to create minimal-size ensembles of urea-denatured ubiquitin (76 aa) that satisfy a very high number of 419 RDC and 253 PRE restraints obtained by NMR (21). The analysis of the ensemble revealed significant (10–20%) populations of a native-like first β-hairpin and α-helix and nonnative α-helical conformations in the C-terminal half. Thus, the urea-denatured state has similarities to ubiquitin's methanol/acid-denatured A-state (28).

Information from SAXS comes at lower resolution but provides important information on the overall shape of a biomolecule in solution (10, 11), which ideally complements the predominantly

Significance

Proteins are the most versatile components of the molecular machinery of life. Synthesized as linear polymers of amino acids, proteins start out in their unfolded state and perform their function either in well-defined folded conformations or as intrinsically disordered proteins (IDPs) lacking tertiary structure. Both for the folding process and the properties of IDPs, a quantitative understanding of the conformational distributions and dynamics of unfolded proteins is thus essential. However, reaching this goal has been challenging owing to the large conformational heterogeneity and rapid dynamics of these systems. Here we combine three of the most powerful biophysical methods available to obtain a comprehensive view of an unfolded protein that would not be available from any of the individual methods.

Author contributions: S.G. and B.S. designed research; M.A., L.D., and A.M.L. performed research; D.N. and J.-r.H. contributed new reagents/analytic tools; M.A., L.D., A.S., D.N., A.M.L., S.G., and B.S. analyzed data; and M.A., A.M.L., S.G., and B.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]M.A. and L.D. contributed equally to this work.

[2]Present address: Department of Chemistry and Interdisciplinary Nanoscience Center, Aarhus University, 8000 Aarhus, Denmark.

[3]To whom correspondence may be addressed. Email: stephan.grzesiek@unibas.ch or schuler@bioc.uzh.ch.

local information from NMR. The SAXS scattering profiles yield the solution-averaged autocorrelation of the molecular electron density, often reported in simplified form as a radius of gyration, R_g.

Single-molecule FRET has recently emerged as an additional method to study unfolded proteins (12, 13). Single-molecule FRET provides intramolecular distance information in the range of tens of angstroms by measuring the Förster transfer between fluorescence donor and acceptor dyes attached to the polypeptide. Because the signal is recorded on single molecules, structural and dynamic heterogeneity can be resolved that is often impossible to detect by ensemble-averaged methods. An example is the investigation of the properties of the unfolded state of a protein in the presence of a folded subpopulation, as it is commonly the case under near-physiological conditions or for IDPs in the presence of ligands (29, 30). The resulting subpopulation-specific transfer efficiencies and fluorescence lifetime distributions can be analyzed in terms of intramolecular distance distributions based on models from polymer physics (3, 31) or combined with molecular simulations (32–34) to quantify global parameters such as mean-squared distances, radius of gyration, or persistence length. In addition to this information relating to equilibrium distances, distance dynamics can be determined on a broad range of timescales with methods such as fluorescence correlation spectroscopy (FCS, $\sim 10^{-8}$ to 10^{-3} s), transfer efficiency histogram analysis ($\sim 10^{-4}$ to 10^{-1} s), microfluidic mixing ($\sim 10^{-3}$ to 10^{2} s) and measurements on immobilized molecules ($\sim 10^{-5}$ to 10^{3} s) (4). Typical reconfiguration times of unfolded proteins as measured by nanosecond-FCS (nsFCS) (30, 35) range between ~20 ns and ~200 ns, a time window that is not accessible by NMR in solution. Thus, the precise long-range and the dynamical information on the >10-ns timescale makes single-molecule FRET highly complementary to both NMR and SAXS.

Despite their obvious complementarity, SAXS, NMR, and single-molecule FRET data on unfolded proteins or IDPs have only rarely been directly combined or compared, and concerns have been raised about the validity of the individual methods for correctly quantifying unfolded-state properties (36, 37). Here we present a direct comparison of the properties of urea-denatured ubiquitin as quantified from all three methods and enhance the description of the ensemble by the FRET-specific single-molecule, time-dependent, and long-range distance information. Compared with previous structural ensembles calculated only from NMR restraints (21), the addition of SAXS restraints (38) did not change significantly either the overall dimensions of the ensemble or its 10–20% secondary structure content similar to ubiquitin's A-state. To obtain a detailed description of the denatured ensemble by single-molecule FRET, seven double-cysteine variants of ubiquitin were created, labeled for FRET, and analyzed. The resulting FRET distance distributions were then compared with the structural ensembles based on NMR and SAXS restraints. The results show good agreement of not only the FRET- and NMR/SAXS ensemble-derived average distances, but also of their distributions. Furthermore, the single-molecule FRET data reveal the reconfiguration times of the unfolded chain in the range of 50–100 ns. Thus, the combined data from all methods provide a detailed and consistent description of the conformational ensemble of urea-denatured ubiquitin at unprecedented spatial and temporal detail.

Results

Structural Ensembles of Urea-Denatured Ubiquitin Calculated from NMR and SAXS Restraints. We had previously calculated structural ensembles of ubiquitin denatured in 8 M urea at pH 2.5 based on experimental NMR observations comprising 419 RDC (20, 39) and 253 PRE restraints determined on eight *S*-(1-oxyl-2,2,5,5-tetramethyl-2,5-dihydro-1H-pyrrol-3-yl)methyl methanesulfonothioate (MTSL)-labeled cysteine variants (21). Ensemble sizes of at least 10 conformers had been sufficient to obtain average back-calculated NMR parameters and radii of gyration that matched the experimental data within the expected error. Here we have extended these calculations by including restraints from the SAXS intensity profile (38) and 71 backbone $^{3}J_{HNH\alpha}$-couplings (40) of urea-denatured ubiquitin as well as increasing the ensemble size to 20. The newly generated ensembles are very similar to the previous ones (21), as can be appreciated from the C^{α}–C^{α} contact map of 400 such ensembles (Fig. 1) that shows highly similar intramolecular interactions corresponding to both native and nonnative structure elements. It is noted that whereas each individual ensemble of 20 conformers matches all experimental data within the error, the different ensembles contain very different structures, because their degrees of freedom largely exceed the number of structural restraints. We have shown previously (21) that the probability for observing individual C^{α}–C^{α} contacts within the total of all structures follows a binomial distribution and that observations of contacts in the low percent range are highly significant for a total of several thousand structures.

In particular, and similar to previous findings, significant contact populations on the order of 10% are observed between residues 6 and 10 as well as 7 and 11 at the location of the native β-turn (Fig. 1). In addition, strong α-helical *i* to *i*+4 contacts exist along the entire peptide sequence, covering both the native helix of residues 22–35 and the nonhelical C-terminal part of the native protein. Thus, the structural ensemble has similarities to ubiquitin's methanol/acid-denatured A-state, in which the first β-hairpin, the α-helix, and a long C-terminal α-helix are formed but move on the nanosecond timescale as independent elements connected by flexible linkers (28). Many further, weaker, native and nonnative long-range contacts are also observed, among them a patch comprising the native contacts between amino acid stretches 18–24 and 50–57. The presence of such long-range

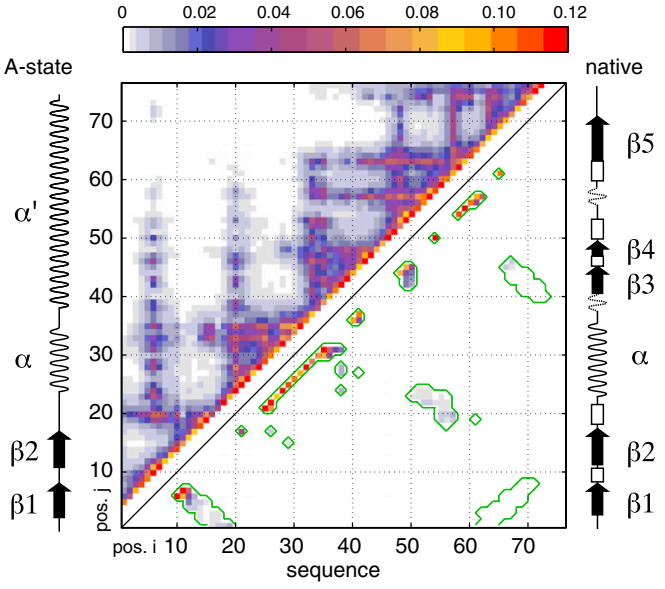

Fig. 1. C^{α}–C^{α} contact map of urea-denatured ubiquitin. Four hundred low-energy structural ensembles of 20 conformers each were generated as described in *Methods* using X-PLOR simulations with restraints derived from RDC, PRE, J-coupling, and SAXS data. Contact probabilities $p(i,j)$ between residue *i* and residue *j* were determined as the total number of contacts observed in all conformers with C^{α}–C^{α} distances smaller than 8 Å divided by the total number of conformers (i.e., 8,000). The contact probabilities $p(i,j)$ are color-coded as indicated by the color bar at the top and plotted versus sequence positions *i* and *j*. The upper left part of the contact map represents all observed contacts; the lower right part shows only contacts also present in the native state of ubiquitin (delimited by green contour lines). Merely for comparison, the secondary structures of native state and A-state ubiquitin are shown schematically at the right and left sides, respectively.

contacts on the order of 1% in the calculated ensemble seems to indicate that the protein samples tertiary interactions of the fully folded native form even in 8 M urea to a detectable extent.

Distance Distributions in the Unfolded Ensemble from Single-Molecule FRET. To probe the distances between different segments of the polypeptide chain in single-molecule FRET experiments and to compare them to the NMR/SAXS ensemble, we created seven double-cysteine variants of ubiquitin with cysteine residues bracketing between 26 and 68 peptide segments. The variants were labeled with Alexa Fluor 488 and Alexa Fluor 594 as FRET donor and acceptor, respectively. We then used confocal single-molecule fluorescence spectroscopy of freely diffusing molecules to obtain transfer efficiency histograms of ubiquitin at different urea concentrations (Fig. 2*A*). The peak at high transfer efficiency corresponds to folded ubiquitin molecules, and the peak at lower transfer efficiencies to unfolded ubiquitin. With increasing concentration of urea, we observe a gradual change in the relative populations of folded and unfolded states and a disappearance of the folded peak at urea concentrations above ∼5.5 M, as expected for urea-induced unfolding of ubiquitin. However, whereas the position of the folded state is unaffected by the addition of urea to within experimental uncertainty in this high transfer efficiency range, the unfolded peak shifts to lower transfer efficiencies, indicating a continuous expansion of the chain with increasing concentration of urea, as previously observed for many unfolded proteins (3, 31, 41–46).

As expected for a disordered chain (30, 43), we also observe a decrease in transfer efficiency for the unfolded state with increasing sequence separation between the two fluorophores at a given urea concentration, as shown in Fig. 2*B* for 8 M urea. For all seven variants of ubiquitin we detect a similar trend for the expansion of the unfolded chain with increasing concentration of urea, but offset corresponding to the length of the segments probed (Fig. 2*C*). To quantify the dimensions of unfolded ubiquitin from the single-molecule FRET data, we converted the mean transfer efficiencies into root-mean-squared distances (RMSDs) using the intrachain distance distributions for a Gaussian chain, previously shown to be a useful approximation for unfolded proteins (32, 41–43, 47). To complement our information on the distances within the unfolded chain of ubiquitin, we additionally performed an analysis of fluorescence intensity decays, which occur on much shorter timescales than unfolded chain reconfiguration (35) and thus provide more direct and robust information about the shape of the underlying distance distribution function (43, 48, 49). For that purpose, we extracted from our single-molecule measurements subpopulation-specific time-correlated single-photon counting histograms of both donor and sensitized acceptor emission for the different variants of ubiquitin at 8 M urea, an example of which is shown in Fig. 3. We then fitted the donor and acceptor signal globally, assuming distributions of transfer rates corresponding to the distribution of donor–acceptor distances for a Gaussian chain (43) and obtained the RMSD values for all variants (see *Methods* for details). The good agreement between the values of RMSD obtained from the mean transfer efficiencies and the fluorescence intensity decays supports the robustness of the analysis (Fig. 3*D*).

Finally, because the single-molecule experiments enable us to quantify unfolded-state expansion over a broad range of denaturant concentrations owing to the separation of folded and unfolded subpopulations in transfer efficiency histograms (Fig. 2), we can use the results to quantify the dependence of unfolded-state dimensions on urea concentration. Concentrated solutions of denaturants such as urea and guanidinium chloride are good solvents for unfolded proteins (50, 51); correspondingly, an increase in the denaturant concentration usually leads to an expansion of the unfolded state (3, 4, 41), as we observe here for ubiquitin (Figs. 2 and 3). This expansion can be expressed in terms of a change in the scaling coefficient, ν, that quantifies the dependence of the RMSD, $\langle r^2 \rangle^{1/2}$, on chain length, N_{aa}, in scaling laws of the form $\langle r^2 \rangle^{1/2} = r_0 N_{aa}^{\nu}$. In bad solvent, ν is expected to be 1/3, in good solvent ∼3/5, and under Θ conditions ∼1/2 (52). Taking advantage of the observation that the prefactor, r_0, for proteins only varies within a narrow range (53) (see *Methods* for details), we estimated the scaling exponent as a function of urea concentration (Fig. 3). Interestingly, we find a continuous increase of ν from ∼0.5 to ∼0.6 between 1.5 M and 9.5 M urea (Fig. 3), indicating that the expansion of unfolded ubiquitin resembles a transition from Θ conditions to the good solvent regime.

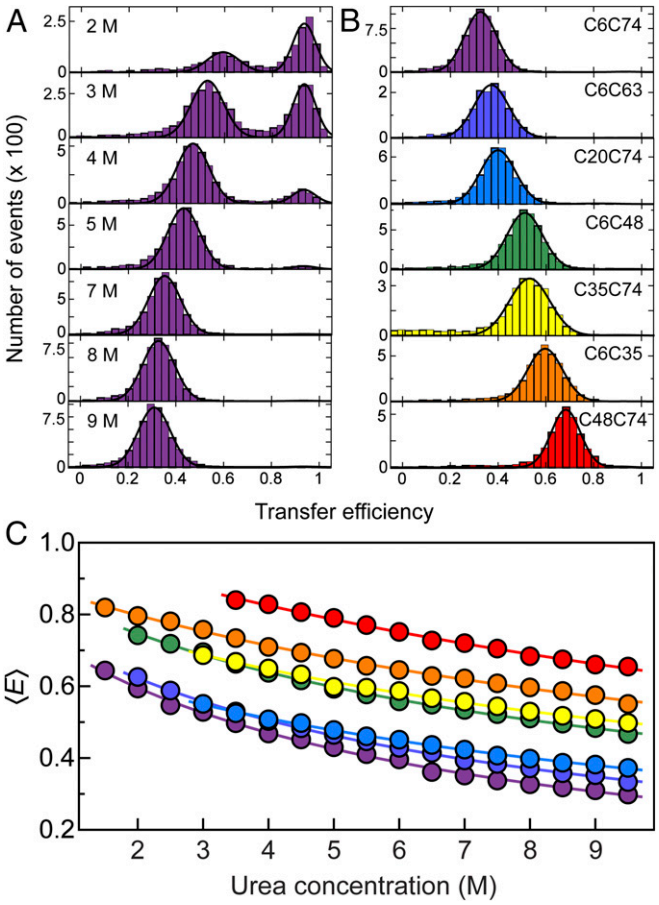

Fig. 2. (*A*) Single-molecule FRET efficiency (E) histograms of the C6C74 variant of ubiquitin at different concentrations of urea, pH 2.5, illustrating the unfolding transition and the unfolded-state expansion. The peak at $E \approx 0.9$ corresponds to folded and the peak at lower E to unfolded molecules. To determine mean transfer efficiencies, $\langle E \rangle$, peaks were fitted with Gaussian peak functions (black lines). (*B*) Histograms at 8 M urea, pH 2.5 for all ubiquitin variants investigated, with the positions of labeled Cys residues indicated for each panel. The color code is the same as in Table 1. (*C*) Dependencies of mean transfer efficiencies of the unfolded subpopulation on the urea concentration for all variants (color code as in *B*). The solid lines are fits with a weak binding model (43, 74) of the form $E(c_D) = E_0 + \Delta E\, Kc_D/(1 + Kc_D)$ for interpolation, where c_D is the denaturant concentration, and K, ΔE, and E_0 are fit parameters.

Comparison of Distances Derived from Single-Molecule FRET and Computed Ensembles. The average distances between amino acids determined by the single-molecule FRET experiments at 8 M urea can be compared directly to distances in the structural ensembles calculated using X-PLOR-NMR. Fig. 4 shows a comparison of the RMSDs between these amino acids in the ensembles versus the respective RMSDs from the

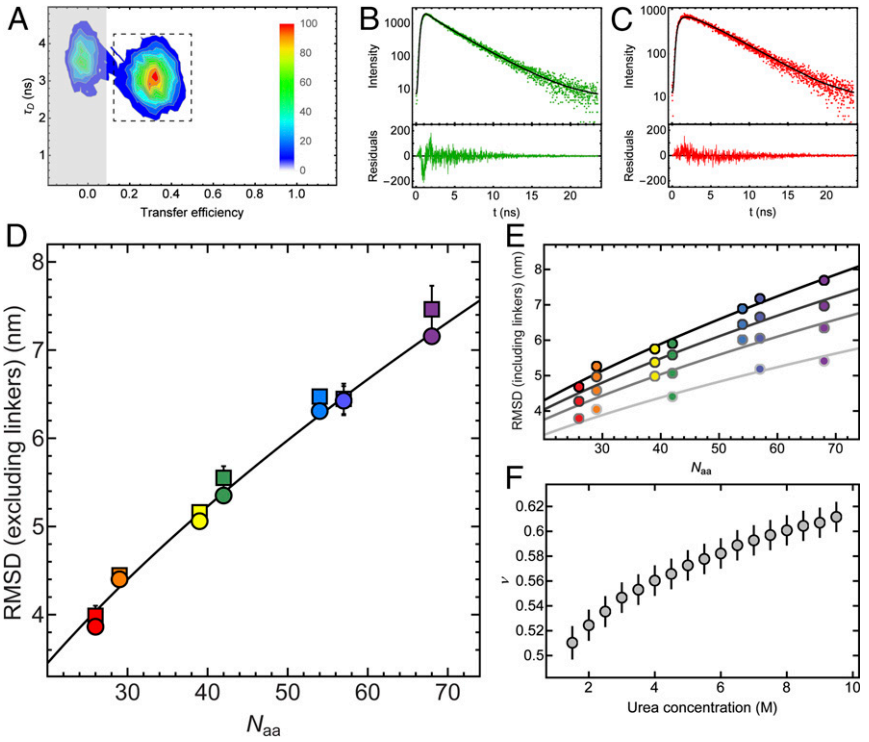

Fig. 3. (*A*) Contour plot of donor fluorescence lifetime, τ_D, versus transfer efficiency for the C6C74 variant in 8 M urea, pH 2.5. Relative amplitudes are indicated as a color scale. For the subpopulation-specific fluorescence lifetime analysis, only the photons from fluorescence bursts in the region indicated by the dashed box were used to exclude the influence of the population containing no active acceptor dye ("donor-only," shaded region). [Note that at lower denaturant concentrations, a population corresponding to folded molecules appears (Fig. 2), whose influence on the lifetime analysis can also be excluded in this way.] (*B* and *C*) The resulting time-correlated single-photon counting histograms of donor (*B*) and acceptor (*C*) emission with residuals below each panel. The donor and acceptor decays were fitted globally based on the distribution of transfer rates resulting from the distance distribution of a Gaussian chain (black lines), with the RMSD as the only free fit parameter apart from the two overall amplitudes for donor and acceptor decays (see *Methods* for details). (*D*) The values of the intrachain RMSDs excluding the dye linkers obtained from the mean transfer efficiencies, either from the numbers of donor and acceptor photons, i.e., the transfer efficiency histograms (Fig. 2), (circles) or from the fluorescence lifetime analysis (squares) for all variants of ubiquitin at 8 M urea are plotted as a function of sequence separation between the labeling sites, N_{aa}. The line corresponds to the equation $\langle r^2\rangle^{1/2} = \sqrt{2 l_p\, b}\; N_{aa}^{\nu}$, where $b = 0.38$ nm is the distance between two C^{α} atoms, and $l_p = 0.43$ nm and $\nu = 0.60$ are the parameters obtained from the global fit (*E*). The colors correspond to the different variants of ubiquitin, with the same color code as in Table 1 and Fig. 2. (*E*) Representative examples of RMSD values from transfer efficiency histograms (including the dye linkers) as a function of N_{aa} for all ubiquitin variants at 2, 4, 6, and 8 M urea [from light gray to black; note that all urea concentrations from 1.5 to 9.5 M (Fig. 2*C*) were used for the analysis] with a global fit to $\langle r^2\rangle^{1/2} = \sqrt{2 l_p\, b}\; N_{aa}^{\nu}$, where l_p was a shared fit parameter for all datasets and ν was allowed to vary with urea concentration. (*F*) The resulting values of ν as a function of urea concentration. Error bars represent standard errors of the fit.

single-molecule FRET measurements. For ensembles generated without any experimental restraints, the RMSDs are systematically larger than those from FRET ($RMSD_{ensemble} \approx 1.40 \cdot RMSD_{FRET}$), but still show a correlation coefficient of $r = 0.996$ with the FRET results. The respective average radius of gyration R_g of the ensemble is 39.3 Å and thus considerably larger than the R_g value from SAXS experiments [28.0 ± 3.5 Å (38)]. Including SAXS or NMR+SAXS restraints reduces R_g to values close to the experiment (SAXS: 28.9 Å; NMR+SAXS: 29.3 Å) and gives excellent agreement with the FRET-derived distances (SAXS: $RMSD_{ensemble} \approx 1.02 \cdot RMSD_{FRET}$, $r = 0.988$; NMR+SAXS: $RMSD_{ensemble} \approx 1.02 \cdot RMSD_{FRET}$, $r = 0.996$). The rms deviation between average distances derived from the $NMR+SAXS_{ensemble}$ and the FRET data is only 1.39 Å. The overall good agreement of distances from the NMR+SAXS-based ensemble and the single-molecule FRET experiments also leads to a good match of the length scaling exponents, ν (NMR+SAXS: 0.61 ± 0.03, FRET: 0.60 ± 0.03; see *Methods* for details). This indicates that the X-PLOR calculations under experimental restraints reproduce the long-range behavior of a random chain, even in the presence of the local, partially ordered structure evident from the C^{α}–C^{α} contact map in Fig. 1.

However, not only the average distances in the ensembles calculated under NMR/SAXS or SAXS restraints closely match the FRET-derived distances, but even the distance distributions within the ensembles are remarkably similar to the distance distributions of a Gaussian chain used for the interpretation of the FRET data (Fig. 5). For the NMR/SAXS-restrained ensemble, the distributions agree almost quantitatively. Any remaining small differences may have several causes: (*i*) genuine specific conformations and local structure induced by the NMR and SAXS data, which are not captured by the simple polymer model used to interpret the FRET observables, (*ii*) incomplete sampling of conformational space in the calculated ensembles, and (*iii*) unknown details of the alignment mechanism for RDCs and the incomplete knowledge of dynamics for PREs, which lead to inaccuracies when incorporating them as quantitative restraints (21). Corresponding to their larger average distances (Fig. 4), the distributions obtained from unrestrained ensemble calculations are much more expanded than those from the restrained ensembles or the FRET-based data (Fig. 5). This behavior is expected for unrestrained X-PLOR calculations, because the remaining nonbonded energy terms contain only repulsive van der Waals terms and a database potential for torsion angles (54) without any further attractive interactions.

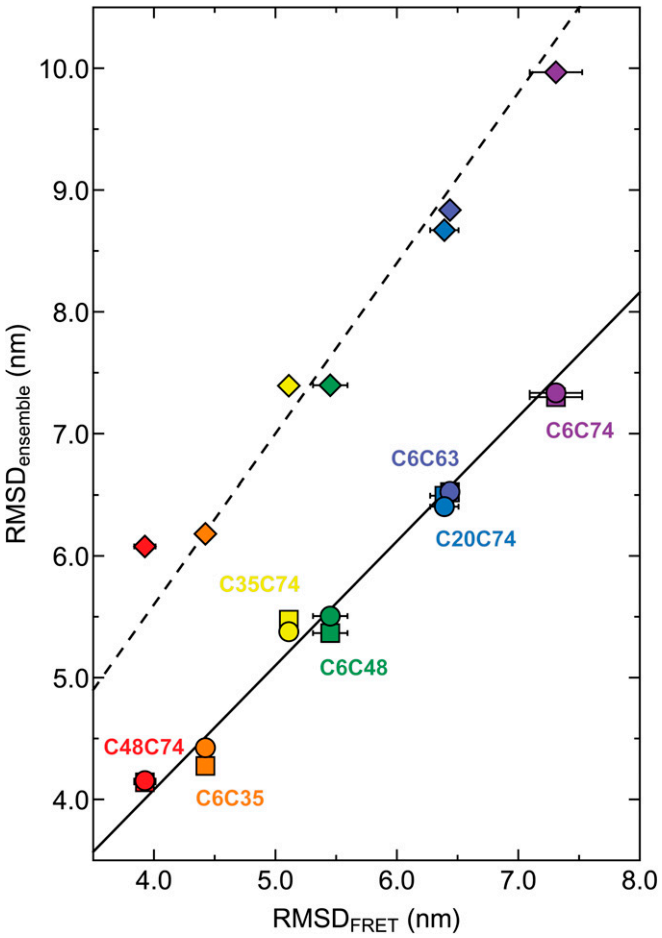

Fig. 4. RMSDs obtained from ensemble calculations versus RMSDs obtained from single-molecule FRET measurements (the average values of RMSDs calculated from mean transfer efficiencies and fluorescence lifetime distribution analysis are corrected for dye linker lengths). Ensemble calculations were carried out with NMR and SAXS restraints (circles), only SAXS restraints (squares), or without any experimental restraints (diamonds), respectively. Color-coding corresponds to respective ubiquitin mutants in Table 1 and Fig. 2. The lines represent least-squares fits to the data of the form $RMSD_{ensemble} = m \cdot RMSD_{FRET}$, with $m = 1.02$ (NMR+SAXS, solid line), 1.02 (SAXS, solid line), 1.40 (no restraints, dashed line).

Reconfiguration Times of the Unfolded-State Ensemble. Single-molecule fluorescence spectroscopy additionally allows us to obtain information about the reconfiguration dynamics of unfolded ubiquitin, which is difficult to access with other methods. For this purpose we used nsFCS to determine the distance dynamics between the donor and acceptor fluorophores attached to the protein (30, 35, 55). Because distance fluctuations in the chain are directly connected to intensity fluctuations of the dyes, the decay of the fluorescence intensity correlation function can be related quantitatively to the relaxation of the distance correlation function (i.e., the reconfiguration time of the polypeptide chain) (35, 56, 57). Examples of fluorescence intensity correlation functions for the C6C74 variant of ubiquitin at 8 M urea are shown in Fig. 6 *A*–*C*, with the characteristic correlated component in the autocorrelation functions and the anticorrelated component in the donor–acceptor cross-correlation in the submicrosecond range (30). The reconfiguration times, τ_r, for all variants of ubiquitin are found to be in the range of 50 ns to 90 ns (Fig. 6*D*), similar to the times observed for other denaturant-unfolded proteins or IDPs (30, 31, 35, 58). We also observe a pronounced dependence of τ_r on the sequence separation of the dyes in the different variants of ubiquitin. We compare our results with the prediction from the Rouse model with internal friction (RIF) (30, 59, 60) (black line in the Fig. 6*D*) and find good agreement between our experimental data and the RIF model with a value for the internal friction time in the range of 10–40 ns. This result indicates a significant contribution of internal friction, slowing down the reconfiguration dynamics of the chain beyond what is expected from solvent friction alone. The value of the internal friction time we observe for ubiquitin in 8 M urea at pH 2.5 is similar in magnitude to that of cold shock protein at ∼4 M GdmCl, pH 7 (30) or spectrin domains between 4 and 8 M GdmCl, pH 7 (58). The molecular origin of internal friction in unfolded proteins is currently unclear; simulations have indicated an important role of dihedral angle rotation (61, 62), but because internal friction is absent in some proteins at very high denaturant concentrations (30, 35), the influence of attractive intrachain interactions (63) and residual structure formation as detected here (Fig. 1) is likely to also contribute.

Discussion

We present here a comparative analysis using single-molecule FRET, NMR, and SAXS to characterize the urea-denatured state of ubiquitin at pH 2.5. Using a previously described approach (21), we generated large sets of ensembles, each containing 20 structures restrained by NMR and SAXS data, and compared them to distance information from single-molecule FRET measurements. Chain dimensions were derived from the single-molecule FRET data on seven donor/acceptor pairs in terms of RMSD, $\langle r^2 \rangle^{1/2}$, using the distance distribution of a simple polymer model as an approximation (32, 41–43, 47). The seven RMSD values show excellent agreement with RMSD values calculated from the atomistic model ensembles that were restrained by NMR and SAXS or SAXS-only data (Fig. 4). Even

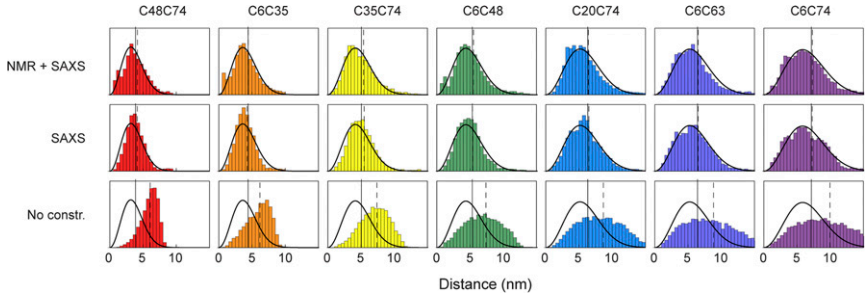

Fig. 5. Comparison of C^{α}–C^{α} distance distributions derived from restrained ensemble calculations and single-molecule FRET. Bars show C^{α}–C^{α} distance distributions from the molecular ensembles calculated under different types of experimental restraints: NMR+SAXS, SAXS, and without experimental constraints (No constr.) (color-coded as in Table 1). Solid black lines depict the distance distributions as obtained from single-molecule FRET measurements for each ubiquitin variant. The vertical black solid and dashed lines indicate the values of RMSDs from FRET measurements and ensemble calculations, respectively.

the distance distributions in the ensembles calculated based on NMR and SAXS restraints agree surprisingly well with the Gaussian chain distance distributions assumed for the interpretation of the FRET data (Fig. 5). Only minor differences exist between the long-range distance distributions of ensembles calculated under NMR+SAXS or SAXS-only constraints, respectively. This observation indicates that the 743 predominantly local NMR restraints can be incorporated into the ensemble structures without disturbing the overall dimensions of the chain. Calculations carried out without experimental restraints show considerably larger average distances and much stronger deviations from the single-molecule FRET data. Thus, the intrinsic X-PLOR force field for the protein bonded and nonbonded interactions (21, 64) does not lead to agreement between the simulated ensembles and the FRET measurements; rather, this agreement has to be enforced by the experimental restraints (Figs. 4 and 5). This is not surprising because the intrinsic X-PLOR force field used contains (besides standard terms for covalent geometry and a conformational database potential for torsion angles) only repulsive van der Waals terms (54), but no attractive interactions such as hydrogen bonding, electrostatics, or hydrophobic effects (65).

The C^{α}–C^{α} contacts of the NMR/SAXS-restrained ensembles are very similar to those of previous ensembles restrained only by NMR RDC and PRE data (21). In particular, we observe prominent contacts of about 15% population between T7 and G10 and between L8 and K11, which indicate the respective population of a β-turn at the location of ubiquitin's native first β-hairpin. This is consistent with earlier findings from ^{15}N and ^{13}C secondary chemical shifts and hydrogen bond scalar couplings of a ∼10–20% population of this β-hairpin even in 8 M urea at pH 2.5 (20). In addition, many further medium to strong (∼5–10%) i to i+4 C^{α}–C^{α} contacts are observed throughout the entire protein sequence (Fig. 1), comprising the native helix in the region of residues 21–31 as well as in the C-terminal half. The location of both the β-hairpin and helical propensities are reminiscent of ubiquitin's A-state induced at low pH and ∼60% methanol (28, 66, 67). However, in the A-state, these secondary structure elements have significantly higher average populations (67). Fig. 1 also shows the presence of residual long-range, native-like contacts in the urea-denatured state at the level of about 1%, which is consistent with a bias required for directing the folding process to the native state (1). The good agreement of the single-molecule FRET-based distance distributions with the NMR/SAXS-restrained ensembles illustrates that such local and residual long-range structure is not inconsistent with the global properties based on simple polymer models (65, 68). We also note that the range of segment lengths accessible to PRE and FRET experiments is complementary. Whereas quantifiable PRE effects for highly unfolded ubiquitin are limited to segments of up to ∼20 residues (21) [corresponding to an average distance of 3.5 nm (Fig. 3*D*)], single-molecule FRET can only probe segments above this length quantitatively given the typical Förster radii of ∼5–6 nm.

Solution NMR provides only limited information on the long-range reconfiguration dynamics of the unfolded protein ensemble. Rotational diffusion of the polypeptide chain makes dynamics unobservable via dipolar, chemical shift anisotropy, or quadrupolar relaxation on timescales longer than the rotational time (69). Recently, Parigi et al. (70) have shown by ^{1}H T_1 relaxation dispersion that a number of IDPs exhibit long-range rotational correlation times in the one-digit nanosecond range, which are comparable to folded proteins of similar size. For urea-denatured ubiquitin, ^{15}N relaxation data (71) indicate an isotropic reorientation of the N-H vectors, which is considerably faster than the rotational correlation time of 4 ns for folded ubiquitin (72). However, slower microsecond motions are evident from chemical exchange line broadening effects for residue T9 in the first β-turn, presumably corresponding to the transient formation of local secondary structure.

Information on long-range dynamics of unfolded proteins from NMR on the timescale of tens of nanoseconds to microseconds has thus been missing for denatured proteins. Here we fill this gap with nsFCS (30, 35), which determines the dynamics of long-range distance fluctuations between the fluorophores attached to the protein. In contrast to NMR relaxation, the FRET-based distance fluctuations observed by nsFCS are not affected by rotational diffusion of the chain if the relative rotational decorrelation of donor and acceptor is faster than the fluorescence lifetime (31, 35). Chain reconfiguration for the polypeptide segments of ∼30 and 70 residues of unfolded ubiquitin occurs on a timescale of 50–90 ns (Fig. 6), despite the residual structure evidenced by NMR. This residual structure may make an important contribution to the internal friction we quantified by analyzing the segment-dependent dynamics with the RIF model (30, 59) (Fig. 6*D*). The timescales derived from correlation spectroscopy agree with the observation that (with the exception of the first β-hairpin) the NMR spectra of unfolded ubiquitin are mostly in fast chemical exchange (39, 71), that is, the interconversion between all conformations is considerably faster than the microsecond timescale of chemical shifts. The NMR observables (i.e., chemical shifts, scalar couplings, RDCs, PREs, and other relaxation rates) are thus averages over all conformations (8).

Owing to these fast global reconfiguration dynamics, the expansion of the unfolded state with increasing concentration of urea is observed to be continuous, as reflected by a gradual shift in transfer efficiency (Fig. 2). This finding is in agreement with previous NMR results (73) showing that most resonances of the unfolded ubiquitin population shift continuously between 8 M and 2 M urea concentration and thus are in fast exchange over the entire urea concentration range. However, below about 3 M urea, a second set of NMR resonances becomes detectable that corresponds to native ubiquitin (73), in accord with the second population detected in the single-molecule FRET efficiency histograms (Fig. 2*A*). Both the NMR data and the single-molecule FRET efficiency histograms show that this folded state is in slow exchange with the unfolded state on a timescale larger than milliseconds. The previous NMR and SAXS data could be described well by a model in which about 30 urea molecules interact as H-bond acceptors with the ubiquitin backbone at 8 M urea, thereby expanding the chain (73). In particular, the continuous chemical shift changes for unfolded ubiquitin in the range of 2 M to 8 M point to a continuous increase in urea binding and chain expansion. This agrees well with the chain expansion observed by the FRET data for increasing amounts of urea, which can be described in terms of an effective weak binding model (74) (Fig. 2*C*). Although direct H-bond interactions with the backbone explain the data well, the exact mechanism of urea-induced denaturation is still under debate (75–78). We note that, based on time-resolved SAXS experiments, unfolded-state collapse of ubiquitin upon jumping from high to low denaturant concentration has been controversial [as for other proteins (36)], with some reports supporting (79, 80) and others contradicting collapse (81). Our results favor the former, but the extent of collapse can of course depend on the solution conditions and the net charge of the chain (53, 82). Resolving this controversy will require further concerted efforts involving a direct comparison of the different methods, either using time-resolved ensemble methods or proteins that remain unfolded even at low denaturant concentrations (83, 84).

In summary, we observe good agreement between distance distributions for unfolded ubiquitin derived from single-molecule FRET and NMR/SAXS-restrained ensembles. The consistency between these orthogonal approaches strongly validates these distance parameters and the underlying computed ensembles. NMR, SAXS and single-molecule FRET data are highly complementary. NMR yields exquisite details on mostly local structural propensities and dynamics, and even residual short- and

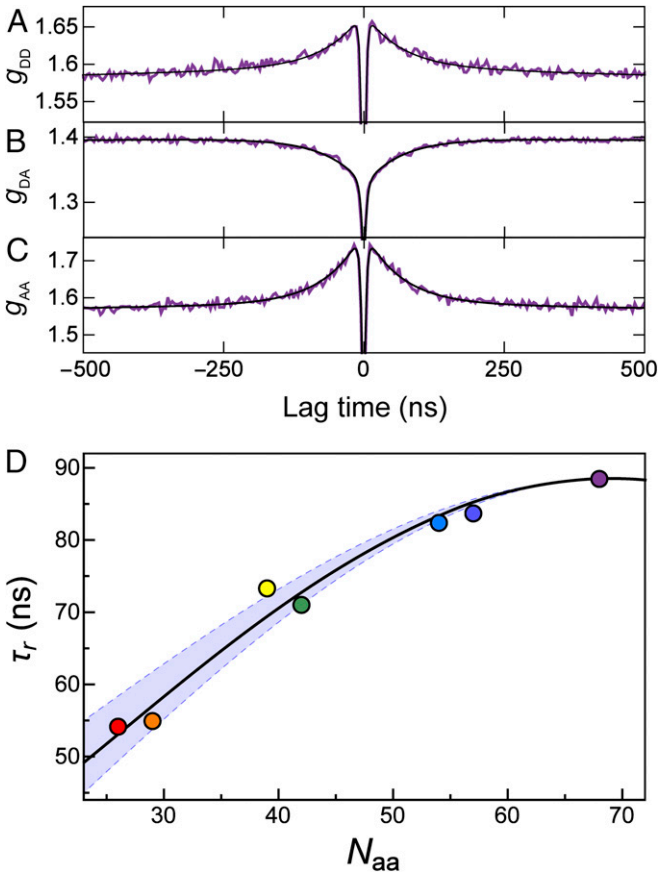

Fig. 6. (*A*–*C*) nsFCS measurements of the donor–acceptor distance dynamics for the C6C74 ubiquitin variant at 8 M urea. The global fit of donor–donor (*A*), donor–acceptor (*B*), and acceptor–acceptor (*C*) correlations is used to determine the reconfiguration time, τ_r, of the unfolded chain. (*D*) The resulting reconfiguration times for all variants as a function of sequence separation show distance dynamics of the unfolded chain in the range of 50–90 ns. The black line represents the fit with a modified RIF model, resulting in an internal friction time of $\tau_i = 25$ ns (see *Methods* for details). The lower and upper bounds of the light blue band are determined by the same fit with internal friction times $\tau_i = 10$ ns and $\tau_i = 40$ ns, respectively. The color code for the variants is the same as in Table 1. Note that the rapid drop in the correlation functions in the low nanosecond range is caused by photon antibunching on the timescale of the excited state lifetime (93) and not related to chain dynamics.

long-range interactions that are only populated in the low percent range within unfolded protein states are detectable. However, the overall dimensions of the unfolded-state ensemble are difficult to define by NMR restraints alone. SAXS data can fill this gap and help to constrain the overall shape of the unfolded-state molecular ensemble. As in the case of most ensemble methods, however, a quantitative analysis is usually limited to conditions where the protein is fully unfolded, because the combined contributions of changes in unfolded-state dimensions and relative populations of folded and unfolded states to the scattering signal is difficult to disentangle. Single-molecule FRET, finally, provides relatively coarse-grained information in the form of average intramolecular distances and distance distributions that are often suitably captured by polymer-physics-based models (31) but cannot resolve the presence of local or transiently formed secondary structure. However, because it is a single-molecule method, conformational heterogeneity can often be resolved, enabling subpopulation-specific analysis; in the case of unfolded proteins, for example, the properties of the unfolded state can be determined even in the presence of folded molecules, and the range of accessible denaturant concentrations can thus be extended compared with ensemble methods. Additionally, segment-specific intramolecular distance information and chain reconfiguration times in the submicrosecond range are accessible, which cannot be obtained by either NMR or small angle scattering methods. Combining this information thus enables us to arrive at the currently most comprehensive description of the structural and dynamic properties of an unfolded protein. Our work thus paves the way for using restraints from all three methods for modeling unfolded states. We expect this type of integrated approach to be ideally suited both for reaching a deeper understanding of the role of unfolded state structure and dynamics in protein folding and for linking the conformational and dynamic properties of IDPs to their cellular functions (61, 62, 85, 86).

Methods

Protein Expression, Purification, and Labeling. Seven double-cysteine variants (K6C-R74C, K6C-K63C, S20C-R74C, K6C-K48C, G35C-R74C, K6C-G35C, and K48C-R74C) of ubiquitin (Table 1) were generated by site-directed mutagenesis from single-cysteine mutants produced previously for MTSL labeling and subsequent PRE measurements (21). The variants were expressed and purified in ^{15}N-labeled form as described before (21) and prepared as stock solutions of 400 μM protein, 50 mM sodium acetate, and 10 mM tris(2-carboxyethyl)phosphine, pH 4.6. The structural integrity of the mutants under these native-state conditions was verified by comparison of their ^{1}H-^{15}N heteronuclear single-quantum correlation NMR spectra to wild-type ubiquitin. Stock solutions were then aliquoted and frozen for later use.

Ubiquitin variants C20C74 and C35C74 were labeled with Alexa Fluor 488 maleimide (Invitrogen) at a dye:protein molar ratio of 0.7:1 for 2 h at room temperature (using ∼200 μL of 0.5 mM protein solutions), and subsequently with Alexa Fluor 594 maleimide (Invitrogen) at a dye:protein molar ratio of 2:1 for 16 h at 4 °C. Unreacted dye was removed by gel filtration with a Superdex Peptide 10/300 GL column (GE Healthcare) in 100 mM sodium phosphate buffer, pH 7.1, and 6 M guanidinium chloride. The resulting sample contained ubiquitin labeled with Alexa Fluor 488 and 594, and ubiquitin doubly labeled with either Alexa Fluor 488 or Alexa Fluor 594, as shown by mass spectrometry. To increase the fraction of FRET-labeled protein for the other variants, purified ubiquitin was first labeled with Alexa Fluor 594 maleimide in 100 mM sodium phosphate buffer, pH 7.1, and 6 M guanidinium chloride, at a dye:protein molar ratio of 0.7:1 for 2 h at room temperature. Unreacted dye was removed by gel filtration using a Superdex Peptide 10/300 GL column (GE Healthcare) in 20 mM Tris·HCl and 4 mM DTT, pH 8.0. Singly labeled ubiquitin was separated from unlabeled and doubly labeled protein with ion exchange chromatography using a Mono Q 5/50 GL column (GE Healthcare). Labeled protein was eluted with a linear gradient of 0–1 M NaCl in 20 mM Tris·HCl and 4 mM DTT, pH 8.0. The fractions containing singly labeled ubiquitin were further labeled with Alexa Fluor 488 after concentration by ultrafiltration (Amicon Ultra, molecular-weight cutoff 3 kDa; Merck Millipore). Twofold excess of Alexa Fluor 488 was reacted with protein singly labeled with Alexa Fluor 594 in 100 mM sodium phosphate buffer, pH 7.1, and 6 M guanidinium chloride for 2 h at room temperature or overnight at 4 °C. After removal of the free unreacted dye, doubly labeled ubiquitin was separated by ion-exchange chromatography, as described above. The resulting sample mainly contained ubiquitin labeled with Alexa Fluor 488 and 594, as confirmed by mass spectrometry. Approximately 100 nmol of unlabeled protein were subjected to the labeling reaction per variant, and ∼10 nmol of pure doubly labeled protein were obtained.

Single-Molecule Fluorescence Spectroscopy. Observations of single-molecule fluorescence were made using a MicroTime 200 confocal microscope equipped with a HydraHarp 400 counting module (PicoQuant) and an Olympus UplanApo 60×/1.20W objective. Alternating excitation of the dyes was achieved using pulsed interleaved excitation (87). The wavelength range used for acceptor excitation was selected with a z582/15 band-pass filter (Chroma) from the emission of a SC-450-4 supercontinuum fiber laser (Fianium) driven at 20 MHz, which triggers the 485-nm pulsed diode laser (LDH-D-C-485; PicoQuant) used for donor excitation. Emitted photons were collected through the microscope objective, focused onto a 100-μm pinhole, and then separated into four channels with a polarizing beam splitter and two dichroic mirrors (585DCXR; Chroma). Photons were additionally filtered by band-pass filters (ET525/50M and HQ650/100; Chroma) before being focused onto one of four single-photon avalanche detectors (SPADs) (Optoelectronics SPCM AQR-15; PerkinElmer or τ-SPADs; PicoQuant).

Table 1. Sequences of the seven ubiquitin variants used in this work

Ubiquitin variant	10	20	30	40	50	60	70	76
K6C-R74C	MQIFVCTLTG	KTITLEVEPS	DTIENVKAKI	QDKEGIPPDQ	QRLIFAGKQL	EDGRTLSDYN	IQKESTLHLV	LRLCGG
K6C-K63C	MQIFVCTLTG	KTITLEVEPS	DTIENVKAKI	QDKEGIPPDQ	QRLIFAGKQL	EDGRTLSDYN	IQCESTLHLV	LRLRGG
S20C-R74C	MQIFVKTLTG	KTITLEVEPC	DTIENVKAKI	QDKEGIPPDQ	QRLIFAGKQL	EDGRTLSDYN	IQKESTLHLV	LRLCGG
K6C-K48C	MQIFVCTLTG	KTITLEVEPS	DTIENVKAKI	QDKEGIPPDQ	QRLIFAGCQL	EDGRTLSDYN	IQKESTLHLV	LRLRGG
G35C-R74C	MQIFVKTLTG	KTITLEVEPS	DTIENVKAKI	QDKECIPPDQ	QRLIFAGKQL	EDGRTLSDYN	IQKESTLHLV	LRLCGG
K6C-G35C	MQIFVCTLTG	KTITLEVEPS	DTIENVKAKI	QDKECIPPDQ	QRLIFAGKQL	EDGRTLSDYN	IQKESTLHLV	LRLRGG
K48C-R74C	MQIFVKTLTG	KTITLEVEPS	DTIENVKAKI	QDKEGIPPDQ	QRLIFAGCQL	EDGRTLSDYN	IQKESTLHLV	LRLCGG

The variant names are colored according to the color code used for the corresponding data throughout the paper. The cysteine residues used for FRET labeling are highlighted in red.

Samples of labeled protein were diluted from concentrations of ~2–5 μM (in 6 M GdmCl and 100 mM sodium phosphate, pH 7.1) to a concentration of ~25–50 pM in measurement buffer (10 mM glycine-HCl, pH 2.5, with concentrations of urea as indicated). Note that despite the low pH, the fluorescence quantum yields of the dyes are essentially unchanged compared with neutral pH (82). We used doubly recrystallized urea for the preparation of single-molecule buffers to minimize fluorescence background often present in commercially available urea. To reduce photobleaching of the chromophores and increase the brightness of the dyes, the photoprotective additive β-mercaptoethanol (143 mM) was added; 0.001% Tween 20 (Pierce) was included to prevent surface adhesion of the protein. All measurements were performed by exciting the donor dye with a laser power of 100 μW (measured at the back aperture of the objective). The power used for directly exciting the acceptor dye was adjusted to match the intensity of donor emission (50–70 μW). Measurements were performed by placing the confocal volume at an axial position 50 μm into the solution relative to the cover slide surface, with an acquisition time of 30 min each.

For the identification of fluorescence bursts caused by proteins diffusing through the confocal volume, we used only the photons detected after donor excitation (i.e., detected in time intervals of ~25 ns following pulses from the 485-nm diode laser). Successive photons separated by interphoton times <100 μs were combined into one fluorescence burst. Only the bursts with a total number of photons >80, detected after donor excitation, were used for further analysis. Transfer efficiencies for each burst were calculated according to $E = n_A/(n_A + n_D)$, where n_D and n_A are the number of photons detected in the donor and acceptor detection channels, respectively, corrected for background, acceptor direct excitation, channel cross-talk, differences in detector efficiencies, and quantum yields of the dyes (88). The changes in refractive index caused by increasing concentrations of urea were measured with a digital Abbe refractometer (Krüss) and were used to recalculate the Förster radius (R_0) for every set of solution conditions. A value of R_0 in the absence of urea of 5.4 nm was used (41). Fluorescence anisotropies (determined for all urea concentrations based on the polarization-sensitive detection in the single-molecule instrument) were between 0.06 and 0.16 for the donor and between 0.04 and 0.13 for the acceptor, indicating rapid orientational averaging of the fluorophores (i.e., $\kappa^2 \approx 2/3$), as usually observed for unfolded proteins (13, 55).

Determining RMSDs from Mean Transfer Efficiencies. The calculation of the root-mean-squared distances of unfolded protein chains from the measured mean transfer efficiencies was performed assuming an interdye distance distribution of a Gaussian chain (32, 43, 47). Experimentally determined mean transfer efficiencies of the unfolded state can be calculated from a dye-to-dye distance distribution, $P(r)$, according to

$$\langle E \rangle = \int_0^l E(r)P(r)dr \Big/ \int_0^l P(r)dr, \quad [1]$$

with $E(r) = 1/(1 + (r/R_0)^6)$, where l is the contour length of the chain. The probability density function for the end-to-end distance of a Gaussian chain is

$$P(r) = 4\pi r^2 \left(\frac{3}{2\pi\langle r^2\rangle}\right)^{3/2} exp\left(-\frac{3r^2}{2\langle r^2\rangle}\right). \quad [2]$$

Given a measured value of $\langle E \rangle$, the RMSD, $\langle r^2 \rangle^{1/2}$, can be calculated numerically. Note that an alternative analysis using Sanchez theory based on a reweighted Flory–Fisk distribution (42, 53, 89) instead of a Gaussian chain distribution resulted in very similar results (ΔRMSD $\leq$ 0.1 nm), supporting the robustness of our results with respect to the distance distribution used, as previously shown for other unfolded proteins and IDPs (46, 47).

The scaling of the RMSD with chain length was analyzed based on the equation $\langle r^2 \rangle^{1/2} = r_0 N_{aa}^{\nu}$ (52), where in the prefactor, $r_0 = \sqrt{2 l_p b}$, l_p is the persistence length, b is the peptide segment length of 0.38 nm, and ν is the length scaling exponent. For the analysis of the RMSD values as a function of the number of peptide segments, N_{aa}, and urea concentration based on the single-molecule FRET data (Fig. 3), l_p was used as a global fit parameter, resulting in l_p = (0.43 ± 0.04) nm. The values of ν given for the FRET-based and the constrained-ensemble-based data at 8 M urea were obtained using this value of l_p; the corresponding uncertainties given for ν reflect the change in ν resulting from the fit when l_p is varied in a range from the lower bound of l_p observed experimentally for proteins in high concentrations of denaturant (90) (l_p = 0.33 nm) to the average value for folded proteins (91) (l_p = 0.53 nm) (53). The RMSDs determined from the analysis of the FRET data include the length contribution from the dye linkers. For comparing the RMSDs of the chain segment probed between FRET and ensemble calculations at 8 M urea (Figs. 4 and 5), the $\langle r^2 \rangle^{1/2}$ values from FRET were corrected for the linker length using $\langle r^2 \rangle^{1/2} = \sqrt{2 l_p b} \cdot N^{0.6}$, where $N = N_{aa} + L$ is the sequence length of the interdye segment, comprising both the number of peptide bonds, N_{aa}, and the contribution from both dye linkers, L. L was previously estimated to be equivalent to an additional nine amino acids (43, 53). A global analysis of the transfer efficiencies for all ubiquitin variants at all urea concentrations investigated (Figs. 2 and 3), with L as a free fit parameter, resulted in L = 9 ± 2 (uncertainty from varying l_p between 0.33 and 0.53 nm), supporting this estimate.

Subpopulation-Selective Fluorescence Lifetime Distribution Analysis. Fluorescence lifetime analysis provides a way of quantifying distance distributions in unfolded proteins orthogonal to the analysis of burst-averaged ratiometric transfer efficiencies (43, 49). For a subpopulation-selective analysis, we use the following procedure. In a first step, for every fluorescence burst, the mean donor fluorescence lifetime was estimated from the average arrival time of the photons in each burst relative to the exciting laser pulse (92). From the resulting 2D histograms of estimated fluorescence lifetime versus ratiometrically determined transfer efficiency (43, 49), we selected the bursts corresponding to the unfolded state and, by using the combined photon counts from all unfolded molecules, obtained subpopulation-specific fluorescence intensity decays (Fig. 3 *A*–*C*). To avoid the common problem of overparameterization and numerical instability in this type of analysis, distance distributions, $P(r)$, of polymer physics-based models can be used, as described previously (43). Here, we used the $P(r)$ of a Gaussian chain (Eq. **3**), which is uniquely defined by $\langle r^2 \rangle^{1/2}$ as the only free parameter (43). The applicability of the Gaussian chain approximation is further supported by the agreement of the corresponding shapes with the distance distributions from the X-PLOR-NIH–derived ensembles (Fig. 5). Additional stability of the fit is achieved by analyzing donor and acceptor decays in a global fit based on the distribution of transfer rates resulting from the distribution of donor–acceptor distances (43, 49). Especially the rise in the acceptor intensity at early times provides a stringent additional constraint. Apart from $\langle r^2 \rangle^{1/2}$, the only other fit parameters were the two overall amplitudes of the donor and acceptor decays. Acceptor direct excitation and cross-talk of donor emission into the acceptor detection channel were taken into account as previously described (43). Intrinsic donor and acceptor fluorescence intensity decay rates were determined independently and constrained in the fits (k_D = 0.275 ± 0.005 ns^{-1} and k_A = 0.250 ± 0.005 ns^{-1} at 8 M urea). Uncertainties in the RMSD values determined by this method (Fig. 3*D*) were obtained by taking into account the uncertainty of the donor fluorescence intensity decay rate in the absence of the acceptor.

nsFCS Measurements. nsFCS measurements were performed in samples with a protein concentration of ~2 nM, with the 485-nm diode laser operating in continuous-wave mode (100 μW) and with typical data acquisition times of 12–16 h to achieve sufficient statistics for interphoton times in the submicrosecond range (35, 55). All measurements were performed in 10 mM glycine-HCl, 8 M urea, pH 2.5, 143 mM β-mercaptoethanol, and 0.001% Tween 20. Autocorrelation curves of acceptor and donor channels and cross-correlation curves between acceptor and donor channels were calculated from the measurements as described previously (35, 55). The data for delay times up to 4 μs were fit with

$$g_{ij}(\tau) = 1 + \frac{1}{\langle N \rangle}\left(1 - c_{AB}e^{-\frac{\tau - t_0}{\tau_{AB}}}\right)\left(1 + c_{CD}e^{-\frac{\tau - t_0}{\tau_{CD}}}\right)\left(1 + c_T e^{-\frac{\tau - t_0}{\tau_T}}\right), \quad [3]$$

where i, $j = A, D$ and $\langle N \rangle$ is the mean number of molecules in the confocal volume. The three multiplicative terms describe the contributions of photon antibunching (93) (*AB*), chain dynamics (*CD*), and triplet blinking of the dyes (*T*) with amplitudes c_{AB}, c_{CD}, and c_T and relaxation times τ_{AB}, τ_{CD}, τ_T, respectively. The three correlation curves from each measurement were fit globally with τ_{CD} and the time origin, t_0, as shared fit parameters. The amplitudes and the relaxation times of the antibunching and triplet components were fit with a free independent decay component for each correlation curve as described previously (30).

The analysis of our experimental data with the RIF model (30, 59, 60) was performed as described previously (30). Briefly, the effect of internal friction on the reconfiguration time of the chain, τ_r, is included in the model as an additional timescale, τ_i, as $\tau_r = \tau_s + \tau_i$, where τ_s is the reconfiguration time of the ideal chain without internal friction. According to the RIF model, τ_s depends on the length of the segment probed owing to the length-dependent relaxation modes in polymer reconfiguration dynamics (30, 59, 94), but τ_i is independent of segment length. Fluorophores and linkers were taken into account explicitly and modeled as additional beads connected to the polymer chain via harmonic springs. We then calculated the reconfiguration times of the unfolded ubiquitin chain as a function of the segment length probed and adjusted the internal friction time, τ_i, to optimize the fit of the model to the experimental data (30). We note that based on this analysis the timescale of the dynamics within the linker segment of the dyes is approximately two orders of magnitude faster than the end-to-end dynamics of the polypeptide chain (30), implying a minor effect on the observed chain dynamics.

Ensemble Calculations. Structural ensembles of urea-denatured ubiquitin were calculated by the program X-PLOR-NIH, version 2.39 (64), using protocols described previously (21) with standard X-PLOR force fields including a conformational database potential for torsion angles and repulsive-only van der Waals terms (54). The ensemble calculations used experimental restraints in different combinations derived from 419 RDCs (nine different types) and 256 PREs published previously (20, 21) that were supplemented by 71 backbone $^3J_{HNH\alpha}$-couplings (40) and SAXS data (38). To speed up the calculation, the SAXS scattering curves were approximated by a 25-point spline fit. Unlike in the earlier work (21), where PRE restraints were introduced in a hypothetical mutant of ubiquitin containing all eight MTSL-labeled cysteines in a single molecule, the PRE restraints were introduced as originating from the C^β atoms or, in the case of glycine, from the C^α atom, of the respective native amino acids in native ubiquitin. Annealing results were independent of initial conformations (extended, folded, or random). Ensemble calculations were carried out using different combinations of restraints: (*i*) NMR and SAXS, (*ii*) SAXS only, and (*iii*) no experimental restraints, which served as a control. Even though an ensemble size of 10 conformers (21) was sufficient to match all experimental NMR and SAXS data within the expected error, larger ensembles led to increased convergence of the intrachain distance distributions, presumably because larger ensembles result in reduced correlations between multiple restraints that need to be matched by an individual conformer. Hence, all calculations were carried out using ensemble sizes of 20 as a compromise between accuracy and computational cost. A total of 800 ensembles with 20 conformers were calculated for each of the restraint sets, out of which the 50% lowest energy ensembles (8,000 total structures) were selected for further statistical analysis.

ACKNOWLEDGMENTS. We thank Marco Rogowski for expert preparation of the ubiquitin cysteine mutants, Dr. Hoi Tik Alvin Leung for very helpful discussions, and Dr. Robert Best for helpful comments on the manuscript. This work was supported by Swiss National Science Foundation Grants 31003A_152839 (to B.S.) and 30-149927 (to S.G.) and postdoctoral fellowship of the Forschungskredit of the University of Zurich FK-13-034 (to M.A.).

1. Zwanzig R, Szabo A, Bagchi B (1992) Levinthal's paradox. *Proc Natl Acad Sci USA* 89(1):20–22.
2. Shortle D (1996) The denatured state (the other half of the folding equation) and its role in protein stability. *FASEB J* 10(1):27–34.
3. Haran G (2012) How, when and why proteins collapse: The relation to folding. *Curr Opin Struct Biol* 22(1):14–20.
4. Schuler B, Hofmann H (2013) Single-molecule spectroscopy of protein folding dynamics–expanding scope and timescales. *Curr Opin Struct Biol* 23(1):36–47.
5. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6(3):197–208.
6. Habchi J, Tompa P, Longhi S, Uversky VN (2014) Introducing protein intrinsic disorder. *Chem Rev* 114(13):6561–6588.
7. Mittag T, Forman-Kay JD (2007) Atomic-level characterization of disordered protein ensembles. *Curr Opin Struct Biol* 17(1):3–14.
8. Meier S, Blackledge M, Grzesiek S (2008) Conformational distributions of unfolded polypeptides from novel NMR techniques. *J Chem Phys* 128(5):052204.
9. Jensen MR, Zweckstetter M, Huang J-R, Blackledge M (2014) Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy. *Chem Rev* 114(13):6632–6660.
10. Millett IS, Doniach S, Plaxco KW (2002) Toward a taxonomy of the denatured state: Small angle scattering studies of unfolded proteins. *Adv Protein Chem* 62:241–262.
11. Bernadó P, Svergun DI (2012) Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Mol Biosyst* 8(1):151–167.
12. Ferreon ACM, Moran CR, Gambin Y, Deniz AA (2010) Single-molecule fluorescence studies of intrinsically disordered proteins. *Single Molecule Tools: Fluorescence Based Approaches, Part A*, ed Walters NG (Academic, New York), Vol 472, pp 179–204.
13. Brucale M, Schuler B, Samorì B (2014) Single-molecule studies of intrinsically disordered proteins. *Chem Rev* 114(6):3281–3317.
14. Neri D, Wider G, Wüthrich K (1992) Complete 15N and 1H NMR assignments for the amino-terminal domain of the phage 434 repressor in the urea-unfolded form. *Proc Natl Acad Sci USA* 89(10):4397–4401.
15. Smith LJ, et al. (1996) Analysis of main chain torsion angles in proteins: Prediction of NMR coupling constants for native and random coil conformations. *J Mol Biol* 255(3):494–506.
16. Gillespie JR, Shortle D (1997) Characterization of long-range structure in the denatured state of staphylococcal nuclease. I. Paramagnetic relaxation enhancement by nitroxide spin labels. *J Mol Biol* 268(1):158–169.
17. Häussinger D, Huang J-R, Grzesiek S (2009) DOTA-M8: An extremely rigid, high-affinity lanthanide chelating tag for PCS NMR spectroscopy. *J Am Chem Soc* 131(41):14761–14767.
18. Shortle D, Ackerman MS (2001) Persistence of native-like topology in a denatured protein in 8 M urea. *Science* 293(5529):487–489.
19. Klein-Seetharaman J, et al. (2002) Long-range interactions within a nonnative protein. *Science* 295(5560):1719–1722.
20. Meier S, Strohmeier M, Blackledge M, Grzesiek S (2007) Direct observation of dipolar couplings and hydrogen bonds across a β-hairpin in 8 M urea. *J Am Chem Soc* 129(4):754–755.
21. Huang J-R, Grzesiek S (2010) Ensemble calculations of unstructured proteins constrained by RDC and PRE data: A case study of urea-denatured ubiquitin. *J Am Chem Soc* 132(2):694–705.
22. Miranker A, Radford SE, Karplus M, Dobson CM (1991) Demonstration by NMR of folding domains in lysozyme. *Nature* 349(6310):633–636.
23. Chen Y, Campbell SL, Dokholyan NV (2007) Deciphering protein dynamics from NMR data using explicit structure sampling and selection. *Biophys J* 93(7):2300–2306.
24. Bernadó P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI (2007) Structural characterization of flexible proteins using small-angle X-ray scattering. *J Am Chem Soc* 129(17):5656–5664.
25. Nodet G, et al. (2009) Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings. *J Am Chem Soc* 131(49):17908–17918.
26. Berlin K, et al. (2013) Recovering a representative conformational ensemble from underdetermined macromolecular structural data. *J Am Chem Soc* 135(44):16595–16609.
27. Schwieters CD, Kuszewski JJ, Clore GM (2006) Using Xplor-NIH for NMR molecular structure determination. *Prog Nucl Mag Res Sp* 48(1):47–62.
28. Brutscher B, Brüschweiler R, Ernst RR (1997) Backbone dynamics and structural characterization of the partially folded A state of ubiquitin by 1H, 13C, and 15N nuclear magnetic resonance spectroscopy. *Biochemistry* 36(42):13043–13053.
29. Ferreon ACM, Gambin Y, Lemke EA, Deniz AA (2009) Interplay of alpha-synuclein binding and conformational switching probed by single-molecule fluorescence. *Proc Natl Acad Sci USA* 106(14):5645–5650.
30. Soranno A, et al. (2012) Quantifying internal friction in unfolded and intrinsically disordered proteins with single-molecule spectroscopy. *Proc Natl Acad Sci USA* 109(44):17800–17806.
31. Schuler B, Soranno A, Hofmann H, Nettels D (2016) Single-molecule FRET spectroscopy and the polymer physics of unfolded and intrinsically disordered proteins. *Annu Rev Biophys* 45:207–231.
32. Nettels D, et al. (2009) Single-molecule spectroscopy of the temperature-induced collapse of unfolded proteins. *Proc Natl Acad Sci USA* 106(49):20740–20745.
33. Nath A, et al. (2012) The conformational ensembles of α-synuclein and tau: Combining single-molecule FRET and simulations. *Biophys J* 103(9):1940–1949.

34. Wuttke R, et al. (2014) Temperature-dependent solvation modulates the dimensions of disordered proteins. *Proc Natl Acad Sci USA* 111(14):5213–5218.
35. Nettels D, Gopich IV, Hoffmann A, Schuler B (2007) Ultrafast dynamics of protein collapse from single-molecule photon statistics. *Proc Natl Acad Sci USA* 104(8): 2655–2660.
36. Yoo TY, et al. (2012) Small-angle X-ray scattering and single-molecule FRET spectroscopy produce highly divergent views of the low-denaturant unfolded state. *J Mol Biol* 418(3–4):226–236.
37. Boomsma W, Ferkinghoff-Borg J, Lindorff-Larsen K (2014) Combining experiments and simulations using the maximum entropy principle. *PLOS Comput Biol* 10(2):e1003406.
38. Gabel F, Jensen MR, Zaccaï G, Blackledge M (2009) Quantitative modelfree analysis of urea binding to unfolded ubiquitin using a combination of small angle X-ray and neutron scattering. *J Am Chem Soc* 131(25):8769–8771.
39. Meier S, Grzesiek S, Blackledge M (2007) Mapping the conformational landscape of urea-denatured ubiquitin using residual dipolar couplings. *J Am Chem Soc* 129(31): 9799–9807.
40. Vajpai N, Gentner M, Huang J-R, Blackledge M, Grzesiek S (2010) Side-chain chi(1) conformations in urea-denatured ubiquitin and protein G from (3)*J* coupling constants and residual dipolar couplings. *J Am Chem Soc* 132(9):3196–3203.
41. Schuler B, Lipman EA, Eaton WA (2002) Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature* 419(6908):743–747.
42. Sherman E, Haran G (2006) Coil-globule transition in the denatured state of a small protein. *Proc Natl Acad Sci USA* 103(31):11539–11543.
43. Hoffmann A, et al. (2007) Mapping protein collapse with single-molecule fluorescence and kinetic synchrotron radiation circular dichroism spectroscopy. *Proc Natl Acad Sci USA* 104(1):105–110.
44. Merchant KA, Best RB, Louis JM, Gopich IV, Eaton WA (2007) Characterizing the unfolded states of proteins using single-molecule FRET spectroscopy and molecular simulations. *Proc Natl Acad Sci USA* 104(5):1528–1533.
45. Ziv G, Thirumalai D, Haran G (2009) Collapse transition in proteins. *Phys Chem Chem Phys* 11(1):83–93.
46. Müller-Späth S, et al. (2010) From the cover: Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc Natl Acad Sci USA* 107(33): 14609–14614.
47. Aznauryan M, Nettels D, Holla A, Hofmann H, Schuler B (2013) Single-molecule spectroscopy of cold denaturation and the temperature-induced collapse of unfolded proteins. *J Am Chem Soc* 135(38):14040–14043.
48. Haas E, Wilchek M, Katchalski-Katzir E, Steinberg IZ (1975) Distribution of end-to-end distances of oligopeptides in solution as estimated by energy transfer. *Proc Natl Acad Sci USA* 72(5):1807–1811.
49. Laurence TA, Kong X, Jäger M, Weiss S (2005) Probing structural heterogeneities and fluctuations of nucleic acids and denatured proteins. *Proc Natl Acad Sci USA* 102(48): 17348–17353.
50. Kauzmann W (1959) Some factors in the interpretation of protein denaturation. *Adv Protein Chem* 14:1–63.
51. Robinson DR, Jencks WP (1963) Effect of denaturing agents of the urea-guanidinium class on the solubility of acetyltetraglycine ethyl ester and related compounds. *J Biol Chem* 238:1558–1560.
52. De Gennes PG (1979) *Scaling Concepts in Polymer Physics* (Cornell Univ Press, Ithaca, NY).
53. Hofmann H, et al. (2012) Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc Natl Acad Sci USA* 109(40):16155–16160.
54. Kuszewski J, Gronenborn AM, Clore GM (1996) Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. *Protein Sci* 5(6):1067–1080.
55. Nettels D, Hoffmann A, Schuler B (2008) Unfolded protein and peptide dynamics investigated with single-molecule FRET and correlation spectroscopy from picoseconds to seconds. *J Phys Chem B* 112(19):6137–6146.
56. Gopich IV, Nettels D, Schuler B, Szabo A (2009) Protein dynamics from single-molecule fluorescence intensity correlation functions. *J Chem Phys* 131(9):095102.
57. Wang ZS, Makarov DE (2003) Nanosecond dynamics of single polypeptide molecules revealed by photoemission statistics of fluorescence resonance energy transfer: A theoretical study. *J Phys Chem B* 107(23):5617–5622.
58. Borgia A, et al. (2012) Localizing internal friction along the reaction coordinate of protein folding by combining ensemble and single-molecule fluorescence spectroscopy. *Nat Commun* 3:1195.
59. Cheng RR, Hawk AT, Makarov DE (2013) Exploring the role of internal friction in the dynamics of unfolded proteins using simple polymer models. *J Chem Phys* 138(7): 074112.
60. Khatri BS, McLeish TCB (2007) Rouse model with internal friction: A coarse grained framework for single biopolymer dynamics. *Macromolecules* 40(18):6770–6777.
61. de Sancho D, Sirur A, Best RB (2014) Molecular origins of internal friction effects on protein-folding rates. *Nat Commun* 5:4307.
62. Echeverria I, Makarov DE, Papoian GA (2014) Concerted dihedral rotations give rise to internal friction in unfolded proteins. *J Am Chem Soc* 136(24):8708–8713.
63. Schulz JC, Schmidt L, Best RB, Dzubiella J, Netz RR (2012) Peptide chain dynamics in light and heavy water: Zooming in on internal friction. *J Am Chem Soc* 134(14): 6273–6279.
64. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. *J Magn Reson* 160(1):65–73.
65. Fitzkee NC, Rose GD (2004) Reassessing random-coil statistics in unfolded proteins. *Proc Natl Acad Sci USA* 101(34):12497–12502.
66. Stockman BJ, Euvrard A, Scahill TA (1993) Heteronuclear three-dimensional NMR spectroscopy of a partially denatured protein: The A-state of human ubiquitin. *J Biomol NMR* 3(3):285–296.
67. Cordier F, Grzesiek S (2004) Quantitative comparison of the hydrogen bond network of A-state and native ubiquitin by hydrogen bond scalar couplings. *Biochemistry* 43(35):11295–11301.
68. Tran HT, Pappu RV (2006) Toward an accurate theoretical framework for describing ensembles for proteins under strongly denaturing conditions. *Biophys J* 91(5): 1868–1886.
69. Palmer AG, 3rd (2004) NMR characterization of the dynamics of biomacromolecules. *Chem Rev* 104(8):3623–3640.
70. Parigi G, et al. (2014) Long-range correlated dynamics in intrinsically disordered proteins. *J Am Chem Soc* 136(46):16201–16209.
71. Wirmer J, Peti W, Schwalbe H (2006) Motional properties of unfolded ubiquitin: A model for a random coil protein. *J Biomol NMR* 35(3):175–186.
72. Tjandra N, Feller SE, Pastor RW, Bax A (1995) Rotational diffusion anisotropy of human ubiquitin from N-15 NMR relaxation. *J Am Chem Soc* 117(50):12562–12566.
73. Huang J-R, Gabel F, Jensen MR, Grzesiek S, Blackledge M (2012) Sequence-specific mapping of the interaction between urea and unfolded ubiquitin from ensemble analysis of NMR and small angle scattering data. *J Am Chem Soc* 134(9):4429–4436.
74. Schellman JA (2002) Fifty years of solvent denaturation. *Biophys Chem* 96(2–3): 91–101.
75. Berteotti A, Barducci A, Parrinello M (2011) Effect of urea on the β-hairpin conformational ensemble and protein denaturation mechanism. *J Am Chem Soc* 133(43): 17200–17206.
76. Hua L, Zhou R, Thirumalai D, Berne BJ (2008) Urea denaturation by stronger dispersion interactions with proteins than water implies a 2-stage unfolding. *Proc Natl Acad Sci USA* 105(44):16928–16933.
77. Stumpe MC, Grubmüller H (2007) Interaction of urea with amino acids: Implications for urea-induced protein denaturation. *J Am Chem Soc* 129(51):16126–16131.
78. Holehouse AS, Garai K, Lyle N, Vitalis A, Pappu RV (2015) Quantitative assessments of the distinct contributions of polypeptide backbone amides versus side chain groups to chain expansion via chemical denaturation. *J Am Chem Soc* 137(8):2984–2995.
79. Larios E, Li JS, Schulten K, Kihara H, Gruebele M (2004) Multiple probes reveal a native-like intermediate during low-temperature refolding of ubiquitin. *J Mol Biol* 340(1):115–125.
80. Qin Z, Ervin J, Larios E, Gruebele M, Kihara H (2002) Formation of a compact structured ensemble without fluorescence signature early during ubiquitin folding. *J Phys Chem B* 106(50):13040–13046.
81. Jacob J, Krantz B, Dothager RS, Thiyagarajan P, Sosnick TR (2004) Early collapse is not an obligate step in protein folding. *J Mol Biol* 338(2):369–382.
82. Hofmann H, Nettels D, Schuler B (2013) Single-molecule spectroscopy of the unexpected collapse of an unfolded protein at low pH. *J Chem Phys* 139(12):121930.
83. Borgia A, et al. (2016) Consistent view of polypeptide chain expansion in chemical denaturants from multiple experimental methods. *J Am Chem Soc*, in press.
84. Zheng W, et al. (2016) Probing the action of chemical denaturant on an intrinsically disordered protein by simulation and experiment. *J Am Chem Soc*, in press.
85. Stanley N, Esteban-Martín S, De Fabritiis G (2015) Progress in studying intrinsically disordered proteins with atomistic simulations. *Prog Biophys Mol Biol* 119(1):47–52.
86. König I, et al. (2015) Single-molecule spectroscopy of protein conformational dynamics in live eukaryotic cells. *Nat Methods* 12(8):773–779.
87. Müller BK, Zaychikov E, Bräuchle C, Lamb DC (2005) Pulsed interleaved excitation. *Biophys J* 89(5):3508–3522.
88. Schuler B (2010) Single molecule spectroscopy of unfolded proteins. *Instrumental Analysis of Intrinsically Disordered Proteins*, eds Uversky VN, Longhi S (Wiley, Hoboken, NJ), pp 371–390.
89. Sanchez IC (1979) Phase-transition behavior of the isolated polymer-chain. *Macromolecules* 12(5):980–988.
90. Kohn JE, et al. (2004) Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc Natl Acad Sci USA* 101(34):12491–12496.
91. Dima RI, Thirumalai D (2004) Asymmetry in the shapes of folded and denatured states of proteins. *J Phys Chem B* 108(21):6564–6570.
92. Eggeling C, et al. (2001) Data registration and selective single-molecule analysis using multi-parameter fluorescence detection. *J Biotechnol* 86(3):163–180.
93. Mets Ü (2001) Antibunching and rotational diffusion in FCS. *Fluorescence Correlation Spectroscopy*, eds Elson ES, Rigler R (Springer, Berlin).
94. Doi M, Edwards SF (1988) *The Theory of Polymer Dynamics* (Oxford Univ Press, New York).